Analysis of Talker Characteristics in Audio-visual Speech Integration

A Senior Honors Thesis

Presented in Partial Fulfillment of the Requirements for graduation with distinction in Speech and Hearing Science in the undergraduate colleges of The Ohio State University

by

Kelly Dietrich

The Ohio State University
June 2008

Project Advisor: Dr. Janet Weisenberger, Department of Speech and Hearing Science

Abstract

Speech perception is commonly thought of as an auditory process, but in actuality it is a multimodal process that integrates both auditory and visual information. In certain situations where auditory information has been compromised, such as due to a hearing impairment or a noisy environment, visual cues help listeners to fill in missing pieces of auditory information during communication. Interestingly, even when both auditory and visual cues are entirely comprehensible alone, both are taken into account during speech perception. McGurk and MacDonald (1976) demonstrated that listeners not only benefit from the addition of visual cues during speech perception in situations where there is a lack of auditory information, but also that speech perception naturally employs audio-visual integration when both cues are available.

Although a growing body of research has demonstrated that listeners integrate auditory and visual information during speech perception, there is a significant degree of variability seen in the audio-visual integration and benefit of listeners. Grant and Seitz (1998) demonstrated that the variability in audio-visual speech integration is, in part, a result of individual listener differences in multimodal integration ability. We suggest that individual characteristics of both the auditory signal and talker might also influence the audio-visual speech integration process (Andrews, 2007; Hungerford, 2007; Huffman, 2007).

Research from our lab has demonstrated a significant amount of variability in the performance of listeners on tasks of degraded auditory-only and audio-visual speech perception. Furthermore, these studies have revealed a significant amount of variability

across different talkers in the degree of integration they elicit. The amount of information in the auditory signal clearly has an effect on audio-visual integration. However, in order to fully understand how different talkers and the varying information in the auditory signal impact audio-visual performance, an analysis of the speech waveform must be performed to directly compare acoustic characteristics with subject performance. The present study conducted a spectrographic analysis of the speech syllables of different talkers used in a previous perception study to evaluate individual acoustic characteristics. Based on behavioral confusion matrices that were made we were able to easily examine possible confusions demonstrated by listeners. Some of the behavioral confusions were easily explained by examining syllable formant tracks, while others were explained by the possibility that noise introduced into the waveform when the stimuli were degraded obscured subtle differences in the voice onset time of some confused syllables. Still other confusions were not easily explained by the analysis completed in the present study. The results of the present study provide the foundation for understanding aspects of the acoustic waveform and talker qualities that are desirable for optimal audio-visual speech integration and might also have implications for the design of future aural rehabilitation programs.

Acknowledgments

I would like to thank my advisor, Dr. Janet M. Weisenberger for providing me the opportunity to work alongside her on this thesis. I have grown personally and professionally through her guidance, support, and insight during this entire process. I would like to thank Natalie Feleppelle for the time, assistance, and guidance she has given me throughout this experience. Finally, I would like to thank my family and friends for their love and constant support.

This project was supported by an ASC Undergraduate Research Scholarship.

Table of Contents

Abstract……2

Acknowledgments……4

Table of Contents……5

Chapter 1: Introduction and Literature Review……..6

Chapter 2: Methods……...18

Chapter 3: Results and Discussion……24

Chapter 4: Summary and Conclusion……...32

Chapter 5: References……...34

List of Figures……36

Figures 1-14……...38

Appendix…...…….51

Chapter 1: Introduction and Literature Review

Speech perception is commonly thought of as auditory only process, but in actuality it is a multimodal process which simultaneously uses both auditory and visual stimuli. Concurrently, this integration process collects information from both auditory and visual stimuli and enables us to combine both stimuli for the understanding of speech. In certain situations where auditory stimuli have been compromised, such as in an individual with a hearing impairment or in a noisy environment, visual stimuli help to make up for missing auditory information. However, McGurk and MacDonald's (1976) study demonstrated that audio-visual integration occurs even when the auditory stimulus alone is perfect.

The McGurk and MacDonald (1976) study was conducted by pairing contrasting audio and visual stimuli. In their study, auditory syllables were dubbed onto a videotape of speakers vocalizing different syllables. The subjects were presented with the auditory syllable [ba] simultaneously with the visual syllable [ga]. The subjects were asked to report the speech sounds they perceived. Two main types of responses occurred. Most often, the response of [da] was reported, a blend in the place of articulation from both stimuli, known as a fusion response. This response indicates that information from both modalities is combined and transformed into an element not presented in either modality (McGurk and MacDonald, 1976). Combination responses were also found when listeners reported the responses of [bagba] or [gaba]. These responses represent a composite of two unmodified elements from each modality (McGurk and MacDonald, 1976). When the same individuals watched untreated film or listened to the syllables without the visual

stimuli they reported the syllables correctly as [ba] and [ga]. Results from the McGurk and MacDonald 1976 study illustrated that even when both auditory and visual cues are entirely comprehensible, both are taken into account when making a perceptual response. Their study concluded that people are influenced by the visual cues they see, a phenomenon now known as the McGurk Effect. The study confirmed that people cannot ignore visual cues because both auditory and visual inputs are taken into account even when both are not necessary.

Auditory Cues for Speech Perception

The acoustic speech signal provides information necessary to identify a speech sound, including cues for place, manner, and voicing in both the temporal and spectral envelopes of the waveform. The place of articulation is cued by formant transitions and describes the location in the oral cavity where articulation takes place during speech production. The places of articulation can include bilabials (with the lips), labiodentals (with the lower lips and upper front teeth), interdentals (with the tongue between the teeth), alveolars (with the tip of the tongue and the alveolar ridge), palatal-alveolars (with the blade of the tongue and the alveolar ridge), palatals (with the tongue and the hard palate), and velars (with the tongue and the soft palate). Manner of articulation is cued through formant intensity and formant frequency changes and describes how the articulators make contact with each other during the production of speech. Stops, fricatives, affricates, liquids, and glides are all ways to describe the manner of articulation. Voicing refers to the state of the vocal folds during the production of a sound. Voicing is cued by voice onset time (VOT), the length of time that passes

between the release of a consonant and the vibration of vocal folds. If the vocal folds are vibrating during production the sound is voiced; if the vocal folds are not vibrating during production the sound is described as voiceless.

Studies have shown that a substantial amount of information can be removed from the speech signal without significantly reducing intelligibility. Studies conducted by Shannon et al., (1995) on auditory speech recognition provided evidence that the speech signal is redundant, containing more information that needed to identify speech sounds. In his study he reduced the spectral information within the speech sound, but preserved the temporal envelope from each of the recorded speech tokens. He replaced the reduced spectral information with band-limited noise while preserving the temporal envelopes from the speech stimuli. In the study, high levels of speech recognition performance could be achieved with only three bands of modulated noise. Syllable identification improved as the number of noise bands modulated by the speech temporal envelope increased. Shannon's study showed that listeners can understand speech sounds when large amounts of information in the spectral waveform are removed. He concluded that waveforms contain a substantial degree of redundant information for speech identification beyond what is necessary for speech understanding.

Remez et al. (1981) studied redundancy in the speech signal by degrading the speech signals into sine-wave speech. The speech signal was reduced to three time-varying sine waves representing natural speech; thus, no traditional acoustic cues were present in the stimuli. Three groups of subjects were presented with different levels of information about the stimuli they would hear. The first group was asked to give their impressions of the stimuli, having been told nothing about the nature of the sounds. The

second group was told they would hear a computer generated sentence and were asked to transcribe the utterance to their best ability. The third group was given extensive information about the stimuli they would hear, including the actual wording. Results of the study showed that the primed listeners could clearly detect and identify the reduced-cue speech, while the naïve listeners did not automatically perceive sinusoid replicas of natural speech as linguistic entities. The results indicated that this reduced structure may not be efficient for spontaneous perception. Remez et al. (1981) concluded that speech perception can endure some absences of acoustic and formant cues only if the natural speech pattern is preserved.

Visual Cues for Speech Perception

While the auditory cues of speech perception place, manner, and voicing are vital pieces to understanding speech signals, visual cues are also contained within speech signals and are important when identifying speech sounds. Unlike the many auditory cues provided by the speech signal, visual cues primarily provide information regarding the place of articulation. These visual cues displayed by a talker consist of movements in the talker's eyes, mouth, or face and sometimes provide a significant amount of information for the listener.

Because of less evident voicing and manner cues, a problem for the listener is created when there is a lack of distinctive characteristics in the visual stimulus when relying on it alone. While some phonemes are easily distinguishable based on visual differences, other phonemes are so similar in visual characteristics that their differences cannot be determined by vision alone. Jackson (1988) describes the phonemes /p, b, m/

as having similar visual characteristics; therefore, listeners cannot distinguish them from each other during speech. Groups of sounds possessing the same visual features, and thus virtually indistinguishable, are known as visemes (Fisher, 1968). Viseme groups usually contain more than one speech sound, all produced with similar movements, creating problems for the listener when relying on the visual cues in speech perception alone. For example, /p, b, m/ comprise a viseme group. They are all bilabial consonants having distinct auditory sounds, but lack a visual difference during production. These sounds have a common place of articulation, but differ in terms of manner and voicing, which are not evident in the visual cues.

Visual cues also create problems due to their high dependency on the talker. Viseme groups can be difficult for speechreaders who rely heavily on the visual components of speech signal. These visemes only allow speechreaders to distinguish between groups of sounds, not individual sounds when determining what others are saying (Jackson, 1988). Speechreading becomes difficult in situations where a person is limited to visual cues only, which are not sufficient to identify the speech sounds being produced. Nitchie uses the term "homophenous" to describe speech sounds that visually appear alike and states that the visual cues are needed, but do not provide enough information to make distinctions (cited in Jackson, 1988). Individual talker differences and the environment in which the sound is produced also have an impact on visual speech perception. Talker variations contribute to significant differences in the viseme groups. Jackson (1988) also found that talkers who were easier to read produced more viseme groups than more difficult talkers who produced fewer viseme groups.

Auditory-Visual Integration Theories

Much of the recent literature has focused on audio-visual speech integration for compromised auditory stimuli. "Audio-visual integration" refers to the processes utilized by receivers to combine information extracted from both auditory and visual stimuli (Grant, 2002). A great deal of individual variability in integration is seen, but visual cues are always helpful for listeners in situations where auditory information may be compromised. Recently, two models have been proposed by researchers to describe audio-visual speech integration. Grant (2002) discusses these two models of speech perception, the Prelabeling Model of Integration and the Fuzzy Logic Model of Perception, in term of their success in predicting listener audio-visual speech integration.

The Prelabeling Model developed, by Braida (as cited by Grant, 2002) is a prediction of auditory-visual recognition made from auditory-only and visual-only confusion matrices (Grant, 2002). This model suggests that all information from both auditory and visual modalities is preserved in a multimodal case, with no interference or biasing from the other modality (Grant and Seitz, 1998). In the Prelabeling model integration is determined by the amount of audio-visual integration the listener produces, the higher the audio-visual scores, the more integration takes place. Grant's assessment of the Prelabeling Model illustrates that some listeners are more efficient at audio-visual integration than others. Grant states that the Prelabeling Model is the best predictor of a listener's audio-visual integration abilities. This model is an excellent source of information for the development of rehabilitative programs which seek to improve audio-visual speech perception.

In contrast, The Fuzzy Logical Model of Perception developed by Massaro (as cited in Grant, 2002) attempts to explain audiovisual integration in which all sources of audio, visual, and audiovisual information are evaluated independently. All stimuli arrive from different sensory channels and are processed prior to their combination and integration. The information is extracted by the listeners and is compared to memory descriptions. All sources are then integrated relative to the memory descriptions and responses are determined based on the degree of support from the stored descriptions (Grant and Seitz, 1998). In this model stimuli integration occurs very late in the process, after visual and auditory inputs are identified. Grant states that the Fuzzy Logic Model is less reliable due to the likelihood of underestimation of a listener's true integration abilities.

Audio-Visual Integration Variability

The processes underlying audio-visual speech integration have been examined in a number of previous studies. Results from these studies have suggested that the process of audio-visual speech integration is different for all listeners. It is proposed that individual differences in listeners, individual talker characteristics, and characteristics of both the auditory and visual speech stimuli all play important roles in the overall speech integration process. Although people receive vast benefit for visual cues in normal and compromised situations, there is a huge degree of variability in the overall amount of speech integration benefit that can be achieved.

Listener Characteristics

Individual integration efficiency explains a substantial proportion of the variability and benefits seen in the audio-visual integration process. Grant and Seitz (1998) stated that the amount of benefit when combining audio and visual speech cues is influenced by the individual's overall ability to integrate. The amount of benefit can differ widely and depends on speech recognition skills, speechreading ability, and language skills (Grant and Seitz 1998). Their study offered other explanations for the differences across subjects on recognition tests due to more obvious reasons such as hearing loss, visual acuity, vocabulary, and the degree of auditory impairment, which all have effects on an individual's audio-visual integration. Grant and Seitz also concluded from their study that better integrators probably pay closer attention to, and are able to extract more information from, the natural associations between the movements of the lips and jaw and the resultant speech sound modulations.

Individual listener characteristics play a large role in the efficiency and individual benefit during the combining of auditory-visual information. Some listeners are more efficient at combining the information from two separate modalities to create a response. Individuals may be better integrators with an auditory-only or visual-only stimulus, but when both are presented and combined their integration may increase or decrease depending on their individual integration efficiency.

Talker Characteristics

Individual talker characteristics can also affect the variability in benefits seen among receivers during the audio-visual integration process. Individual talkers are

shown to produce different outcomes for listeners in the integration process. Studies demonstrate that some individuals may be better integrative talkers than others, affecting the listener's ability to integrate. Facial cues and movements given off by the talker may also have an effect on a listener's integration process. Information from a clear talker enables the listeners to integrate easily if they are efficient in combining the information received from both modalities.

A pilot study in our laboratory focused specifically on talker differences. Andrews (2007) examined various talker characteristics to determine which characteristics produce optimal auditory-visual integration. The study evaluated fourteen talkers for speech intelligibility when producing isolated syllables. As each talker produced a set of single-syllable speech tokens they were video recorded and their voices were recorded through a microphone directly into a computer. The auditory samples were degraded using MATLAB by swapping the temporal envelopes of the speech waveforms and the broadband noise, resulting in an auditory signal containing a preserved speech envelope with the fine structure removed (Shannon et al., 1995). The degraded auditory signals were dubbed onto visual samples of a talker producing a single syllable speech token. The listeners were presented with the stimuli under three conditions; 1) degraded auditory-only, 2) visual-only, 3) degraded auditory-visual. Listeners were asked to identify the speech syllable presented and examinations of the performance correlations were completed. Results of Andrews (2007) study showed wide variability in talker intelligibility. Surprisingly, overall intelligibility in the auditory condition was not a strong predictor of performance in the auditory-visual condition. These results leave many unanswered questions. For example, why do two talkers with

the same overall auditory intelligibility produce very different levels of audio-visual speech benefit?

Auditory and Visual Speech Stimuli Characteristics

Variability in benefit levels found throughout the audio-visual integration process is also observed as a function of the acoustic characteristics of the auditory signal. Variability in the auditory signal is illustrated in the production of speech through acoustic differences in spoken words.

Two pilot studies previously completed in our laboratory (Huffman, 2007 and Hungerford, 2007) examined how altering characteristics of the auditory signal impacted auditory-visual integration. Through isolating and systematically removing information from the auditory signal, they studied the response patterns that were altered when visual stimuli were added. In their studies the auditory stimuli were degraded using a method similar to Shannon et al. (1995); auditory syllables were reduced to a waveform composed of a broadband noise fine structure that is modulated by the temporal envelope of the original speech stimulus recording. Each degraded speech stimulus was then filtered into two, four, six, or eight spectral bands, effectively reducing the speech signal information. Both studies found that some of the place, manner, and voicing cues may be lost due to the noise fine structure of the speech signals. Results of their studies indicate that listeners perform better when more auditory information is available; however, removing information from the auditory stimulus does not necessarily affect the degree of integration achieved.

There are a number of unanswered questions regarding why talkers elicit different amounts of integration with auditory and visual stimuli. What are still unknown are the acoustic characteristics in the speech waveform that facilitate listener integration. Does the amount of redundancy between visual and auditory aspects of a speech waveform result in improved integration for a listener? Are there certain individual talker characteristics that promote a listener's ability to integrate stimuli? Finally, does overall intelligibility of auditory information predict more efficient integration?

Analysis of Talker Characteristics in Audio-visual Speech Integration

The present study further explored studies completed by Huffman (2007), Andrews (2007), and Hungerford (2007) by examining the acoustic characteristics of stimulus words produced by the talkers in these studies and comparing these to behavioral confusion matrices for these talkers. This study compared the best and worst talkers from their studies in two channel, four channel, and undegraded auditory conditions. Auditory speech tokens from a subset of the fourteen talkers previously studied were selected from the behavioral confusion matrices generated by these talkers. The present study compared the waveform analysis to behavioral results from Huffman (2006) and Hungerford (2007) to determine whether acoustic characteristics of a given talker can predict the integration process. The tokens were analyzed acoustically using spectrographic analysis computer software to evaluate F2 formant transitions, manner characteristics, and voicing components, primarily in initial stop consonants. From the results we may be able to determine whether particular acoustic characteristics facilitate integration. The results of this study should provide further insights into the mechanisms

governing auditory-visual speech perception and may also have implications for the design of future aural rehabilitation programs.

Chapter 2: Method

The present study used stimuli and behavioral confusion matrices from Huffman (2007). In this section we describe both the methods used by Huffman (2007) to create and present stimuli, and the procedure used in the present study to analyze the acoustic tokens.

Huffman (2007) Study

Participants

Participants in the present study included talkers and listeners who originally participated in Huffman's (2007) study. The talkers consisted of three female and two male participants with ages ranging from 20 to 23, who each produced a set of eight syllabic stimuli that were recorded by a video camera. All of the talkers were undergraduate/graduate university students and reported having normal hearing and normal or corrected vision. The listeners consisted of eight female and two male participants ranging in from ages 17 to 22. Three of the ten listeners were undergraduate university students in the Speech and Hearing Science major. All ten listeners reported having normal hearing and normal or corrected vision.

Stimuli

A set of eight CVC syllables were used as the stimulus words in Huffman's (2007) study to test auditory information in audio-visual integration. For each of the

conditions, auditory-only, visual-only, and auditory-visual, the same stimuli were administered: bat, cat, gat, mat, pat, sat, tat, and zat. These eight stimulus words were chosen for their ability to satisfy the following conditions:

1. Pairs of the stimuli were minimal pairs, differing only by the initial consonant
2. All stimuli were accompanied by the vowel /ae/, which does not exhibit lip rounding or lip extension.
3. Multiple stimuli were used in each category of articulation, consisting of: place (bilabial, alveolar), manner (stop, fricative, nasal), and voicing (voiced, unvoiced).
4. All stimuli were presented without a carrier phrase (citation-style)
5. Stimuli were known to elicit McGurk-like responses

Auditory Signal Degrading

Each of the talkers produced a set of eight monosyllabic stimuli words five times each. Their voices were recorded through a microphone directly into a computer, using the software program Video Explosion Deluxe, which stored the files in .wav format. The auditory files were converted into degraded auditory speech samples using a MATLAB subroutine, created by Bertrand Delgutte (Smith, Oxenham & Delgutte, 2002). Each speech signal was filtered into two and four spectral bands providing equal spacing in basilar membrane distance. The cutoff frequencies for the two spectral bands were 80 Hz, 1,877 Hz, and 19,200 Hz. The cutoff frequencies for the four spectral bands were 80 Hz, 518 Hz, 1,877 Hz, 6,097 Hz, and 19,200 Hz. The syllables were then reduced to a waveform composed of broadband noise fine structure that is modulated by the temporal

envelope of the original speech stimulus recording. The waveform containing the noise fine structure and the temporal envelope cues of the original speech signal is preserved, while the other waveform is discarded. The resulting auditory stimuli in this experiment were degraded in a manner similar to those created by Shannon et al. (1995).

Digital Video Editing

Visual stimuli were obtained for the study by recording two male and three female talkers with a digital video camera, who repeated a list of eight words five times each. The auditory and visual stimuli were downloaded into a computer program called Video Explosion Deluxe. This program was used to edit audio and visual clips. The program enabled the degraded .wav files produced in MATLAB 5.3 to be dubbed onto any visual clip. Randomized lists were made and visual talker clips were paired randomly with auditory clips of tokens of the appropriate syllable. From the lists, videos featuring sixty stimulus clips were created in Video Explosion Deluxe. Sonic MY DVD was the software program used to burn the individual videos created in Video Explosion Deluxe.

Testing Procedure

The present study tested all observers in the basement laboratory room of Pressey Hall, part of The Ohio State University's Speech and Hearing Department. The room provided a space for testing, including a well-lit quiet environment and sound-attenuating booths. A chair was placed on the back wall of the booth giving the observers the ability to see through the double-glass window in the sound-attenuating booth. TDH 39

headphones were used for auditory stimulus presentation. Visual stimuli were presented on a twenty-inch video monitor placed directly outside the booth's glass window, at a distance of about one meter from the observer's face. Visual stimuli were presented on the monitor such that the talker's face was approximately seven inches in height.

Each observer was given a set of instructions in which they were told they would be tested under three randomized conditions: degraded auditory-alone, visual-alone, and degraded audio plus visual. They were instructed that during each condition they would be presented with sixty stimulus words and that a verbal response was needed after each stimulus word was presented. The observers were told that all sixty of the stimuli were phonemes that ended in "at". They were also told that any consonant or combination of consonants could form the combinations and that they may or may not exist in the English language.

Each observer was tested in a quiet, sound attenuating booth. The observers were tested under the three stimulus conditions; in each condition sixty stimulus words were presented via DVDs. Five talkers were presented, each in 2-channel, 4-channel, 6-channel, and 8-channel conditions. The stimuli were randomly presented to the observers who provided a verbal response to each stimulus. All observer responses were recorded by the examiner.

The Present Study (2008)

Analysis of Talker Characteristics in Audio-visual Speech Integration

Speech Integration

For this study two talkers who previously participated in Huffman's (2007) study were chosen for acoustical analysis of talker characteristics. Talker LG, the best talker, and talker JK, the worst talker from the Huffman (2007) study were chosen in terms of overall percent correct intelligibility. This study analyzed five tokens of each of the eight stimulus words from each talker in the 2-channel and 4-channel conditions.

Acoustic Analysis

A spectrographic analysis computer program called TF32, a time-frequency analysis for 32-bit Windows, was used to acoustically analyze the speech waveforms for both talkers' syllable productions. All eight syllables, each produced five times by each talker, were analyzed. The degraded stimuli for each of the eight words in the two channel, four channel, and undegraded conditions were imported into the program. Using TF32 we were able to change time-frequency settings that enabled clear analysis of the formant transitions and manner characteristics. The bandwidth was set to 450 Hz within the lowest frequency range, approximately around 4,500 kHz. While using the program we were also able to modify the parameters of the noise floor (dB) and dynamic range (dB), which we manipulated to provide a more visible representation of the auditory stimulus being analyzed.

Data Collection

Endpoint values were extracted at five points along the wave; 0, 25, 50, 75, 100 percent of the waveform duration for each of the formants (F1, F2, F3). The averages of these values for each stimulus word in each condition were used to generate formant frequency (Hz) plots over the duration waveform in percent. These plots enabled visual inspection of formant tracks for each talker in each condition for each of the eight stimulus words.

Chapter 3: Results and Discussion

Behavioral Confusion Matrices

Behavioral confusion matrices were constructed to aid in selecting stimuli to evaluate by acoustic analysis. Figures 1-4 show the confusion matrices for the two selected talkers, LG and JK, for the stimuli presented in the 2-channel and 4-channel conditions. As mentioned, talker LG produced overall highly intelligible stimuli, whereas talker JK produced relatively unintelligible stimuli based on the results in Huffman (2007). However, it is of interest to evaluate performance for specific stimuli and to analyze the perceptual confusions, to determine how variations in the acoustic waveforms might contribute to perceptual confusions. Specifically, stimuli that showed both relatively low percent correct and substantial concentration of confusions with a few other stimuli were selected. The present analysis focused primarily on the analysis of F2 transitions for the stop consonant syllables in the 2-channel and 4-channel conditions.

As a first effort in determining cues used by listeners to identify these stimuli individually, speech tokens were analyzed using the computer software program TF32, which allowed us to follow formant tracks over the duration of the waveform.

2-Channel Condition

Visual analysis of formant tracks for stimuli from the 2-channel confusion matrices allowed us to explain some of the confusions. However, others were surprising. Evaluation of the stimuli was based on the assessment of the F2 transitions in the

stimulus tokens. We focused on the stop consonants in the 2-channel condition. In Figures 1 and 2 behavioral confusion matrices are shown for the 2-channel conditions.

In Figure 1, it can be seen that talker JK's auditory stimulus "pat" was confused with "cat" twenty-six percent of the time. Figure 5 depicts formant tracks for "cat" and "pat" stimuli produced by this talker. The stimulus "pat" shows a rising F2 transition, which is expected to be found in an undegraded stimulus. "Cat" is expected to have a decreasing F2 transition. This is not apparent in the figure, suggesting that "cat" was confused for "pat" due to the anomaly in the F2 transition.

The stimulus "gat," characterized by a decrease in F2, was often confused for "bat," which is characterized by a rising F2 transition. Figure 6 shows "gat" and "bat" formant tracks for talked JK in the 2-channel condition. Acoustic analysis showed that "gat" might be confused for "bat," given that the "gat" token shows a rising F2, resembling a "bat" transition.

Also in the 2-channel condition, the stimulus "tat" was often confused with "cat" and "pat." The confusion of "tat" with the word "pat" by observers was explainable through the acoustic analysis. Figure 7 shows formant tracks for "tat," "pat," and "cat" for talker JK in the 2-channel condition. While "tat" is typically characterized by a slight falling F2 transition, "pat" is characterized by a rising F2 transition. In this case, the figures show the F2 transition for "tat" as rising like that for "pat," probably accounting for the confusion of "tat" with "pat." "Tat" was also confused with "cat" forty-eight percent of the time. However, the occurrence of confusion between "tat" and "cat" was not easily explained by the F2 transitions.

Figure 2 depicts the behavioral confusion matrices for talker LG in the 2-channel condition. Although the stimulus syllable "gat" was correct fifty-seven percent of the time, talker LG still created some confusion for observers with the response of "cat." These two syllables are shown in Figure 8. "Gat" is a voiced syllable characterized by a falling F2 transition and was confused for "cat," a voiceless syllable characterized also by a falling F2 transition. Both syllables are velar stops and are only differentiated by their voicing component. We might expect this confusion between "gat" and "cat" due to the noise added to the signal during the 2-channel degrading process. The voiced component in the word "gat" therefore was perceived as the word "cat," a voiceless velar stop.

LG also produced listener confusions with the stimulus word "tat." "Tat," characterized by a slight decrease in the F2 transition, was confused almost fifty percent of the time with "cat," which is characterized by a sharp decrease in the F2 transition. Figure 9 shows talker LG's productions of these two stimuli. This confusion we can explain based on the "cat" syllable's slight decrease in the F2, giving observers reason to believe it was "tat." "Tat" is also confused with the syllable "pat" twenty-three percent of the time, which cannot be explained from our analysis of the F2 transition waveform acoustics.

4-Channel Condition

In the 4-channel condition talkers JK and LG produced some easily explained confusions, but also others that were more surprising. Evaluations of the stimuli were again based on the assessment of F2 transitions. Figures 3 and 4 show the behavioral confusion matrices for talkers JK and LG in the 4-channel condition.

Figure 3 shows talker JK's behavioral confusion matrix in the 4-channel degraded condition. The behavioral confusion matrix shows that "gat" was confused with the word "bat" 21% of the time, and with "pat" 16% of the time. This confusion is somewhat difficult to explain. The stimulus word "gat" shows a sharp decrease in the F2 transition, which would be expected in the undegraded state. Thus, it is a bit surprising that this stimulus was not identified correctly a higher percentage of the time. One possible explanation may be found in the relative ambiguity of the stimuli for "bat" and "pat" produced by JK in the 4-channel condition. Both "bat" and "pat," expected to be characterized by a rising F2 transition, instead show a flat to slightly falling F2 transition. Their presence in the stimulus set may have influenced observer response to other stimuli in the set, such as "gat."

The response of "dat" was responded for the stimulus "gat" twenty-nine percent of the time. While we know that "dat" is typically characterized as having a slight F2 decrease, "dat" was not used as a stimulus word in the present study. We are not able to explain this confusion without knowledge of how "dat" might have been represented in the 4-channel condition for talker JK.

Talker JK produced confusions for the stimulus "tat" with a response of "pat" almost fifty percent of the time. Figure 11 shows the formant tracks for stimulus "pat" and "tat" for talker JK in the 4-channel condition. While in an undegraded state "pat" is characterized as having a rising F2 transition, in the 4-channel degraded condition it shows a flat F2 transition, probably accounting for the confusion with "tat," characterized by a slight decrease in the F2 transition.

Figure 4 shows the behavioral confusion matrices produced from observer responses for talker LG. The stimulus word "gat" was correctly responded to sixty-six percent of the time, but was still confused by observers. "Gat," a voiced velar stop, was confused with "cat," a voiceless velar stop. This result is not surprising due to the possible interference with the Voice Onset Time difference that may have been produced by the noise in the 4-channel degraded waveform.

Although observers responded correctly to the stimulus word "cat" sixty-five percent of the time it was also confused twenty-one percent of the time in the observer response of "pat." Figure 12 depicts formant tracks for "cat," "pat," and "tat" by talker LG in the 4-channel condition. "Cat" is characterized by a decreasing F2 transition and "pat" is characterized as a drastic increasing F2 transition. In Figure 12, "pat" does not increase drastically in F2, possibly making it seem more like the syllable "cat" to observers. In the 4-channel condition, "tat" was also confused with "pat" and "cat," together almost twenty-six percent of the time. The "pat" response by observers is explainable by the slight decrease in F2 for the stimulus "pat," which resembles "tat" in the F2 transition created by the 4-channel degrading. The response of "cat" is also reasonable due to the similarity in the decrease of both "tat" and "cat" in the F2 transitions of the formant tracks.

As a first look the present study focused on the stop consonant similarities and differences in the F2 transitions of the waveforms that can be seen through the spectral information provided by the computer software program, TF32. Additional analysis needs to be performed for fricative stimuli, where other spectral factors might be appropriate to examine.

Audio-visual Integration

What does all this information have to do with integration? Talkers LG and JK both show visual percent correct at about thirty percent as indicated in Figure 13. As also shown in this figure, JK shows slightly more audio-visual integration, defined as the improvement from auditory only to auditory plus visual, about ten percent for JK and about five percent for talker LG. This information suggests that some auditory information from LG may be more redundant with the visual information presented and for talker JK auditory and visual information may be more independent and less redundant.

Overall intelligibility of auditory information does not predict integration. Talker LG shows better performance in the auditory alone condition, but overall shows less integration of information. This analysis supports the idea that information integration may be a process independent of auditory and visual stimuli alone performance for specific talkers.

To evaluate this further, confusion matrices for the auditory-visual presentations for talker JK and LG were examined. When adding visual information to the presentation of a stimulus the listeners should produce an overall improvement in the number of correct responses. By adding visual information to an auditory stimulus the listeners were better able to distinguish cues for place of articulation and also some manner characteristics. When presentation in the auditory-visual condition took place we predicted more correct responses than shown in auditory-only presentation. We also predicted to see listeners produce confusions with stimuli that have similar place

characteristics. An example of this would be confusions with bilabial sounds, such as mat, pat, and bat where an obstruction occurs between the lips.

Figure 1 and figure 15 illustrate the behavioral confusion matrices for talker JK in the 2-channel auditory-only and auditory-visual conditions. Comparing these figures there is an average increase of 15.88% in the number of correct responses from the auditory-only condition to the auditory-visual condition. As illustrated in figure 1, the listener responses to stimulus presentation in the auditory-only condition are seen across the matrix. When the visual cues were presented in the auditory-visual condition the listener confusions were usually found among words with similar place of articulation information, as shown in figure 15. This result illustrates that the added visual information provided improvement for listeners to determine the stimulus presented based on place of articulation cues. When the stimulus "pat" was presented with added visual cues a 49% increase in correct responses was seen based on the number of correct responses in the auditory-only condition. In the auditory-visual condition the stimulus word "gat" had an increased correct response of 23% and "cat" had an increase by 15% in the correct responses, while there were still a spread of responses. The stimuli "gat" and "cat" have less visible place of articulation cues consequently making the added visual presentation less significant information for the listeners.

Figure 2 and figure 16 illustrate the behavioral confusion matrices for talker LG in the 2-channel auditory-only and auditory-visual conditions. These figures show an average increase of about 15.5% when the visual information was added to the presentation of an auditory-only stimulus. There was a 17% increase in correct responses when listeners were presented with "mat" in the auditory-visual condition versus the

auditory-only condition. When the stimulus "sat" was presented in the auditory-only condition there was a 15% correct response rate and when it was presented in the auditory-visual condition there was a 52% increase in correct response rates reaching 67% correct. The place information presented for "sat" in the auditory-visual condition eliminated the incorrect confusions with the bilabials that listeners had previously made in the auditory-only condition.

Grant and Seitz (1998) also found a great deal of individual differences across listeners in the amount of audio-visual benefit achieved. The results suggest that perfect integration may not be associated with high overall auditory, visual, or audio-visual performance; rather that they may be independent predictors. The results suggest that overall audio-visual intelligibility is greatly influenced by the amount and type of unimodal cues available. A subject with poor hearing might be predicted to have a low audio-visual performance score, but under these conditions may integrate the auditory and visual cues in a nearly optimal manner. The amount of benefit from combining auditory and visual in everyday situations may also be dependent on a number of factors, including the degree of auditory impairment, speechreading ability, and language skills. Remaining differences might be attributable to differing efficiency in the operation of a perceptual process that integrates auditory and visual speech information.

Chapter 4: Summary and Conclusion

Overall, the present study provided the foundation for understanding acoustic patterns and talker qualities found in acoustic waveforms that are desirable for optimal audio-visual speech integration. Various perceptual confusions demonstrated by listeners in the present study were easily explained by examining the formant tracks of the confused speech stimuli. Other perceptual confusions were explained by addressing differences in the voice onset time possibly created during the syllable degrading process where noise may have been introduced into the speech waveform obscuring subtle differences, resulting in perceptual changes in the voice onset time of the confused syllables.

However, the present study also encountered behavioral confusions from listener responses that were not easily explained by the spectral analysis that we performed. The present study was limited in some ways due to its focus on particular aspects of the speech waveform, such as the F2 transitions in the formant tracks, which only accounted for confusions in some of the selected stimuli. The present study also restricted its primary focus to place characteristics of stop consonants.

The present study verified that the amount of information provided by the acoustic and visual stimuli does impact on the process of audio-visual integration. The amount of redundancy in the auditory and visual information presented to a listener plays a role in the amount of integration and benefit for that listener. Less redundant and more independent auditory and visual information seems to have a greater positive effect on the audio-visual integration process and benefit.

Overall, the present study was just a first look at the question of whether particular acoustic characteristics facilitate the highest amount of audio-visual integration and produce the most benefit to listeners. Future studies might examine additional aspects of the acoustic waveform to determine if other cues might explain more of the perceptual confusions. Further studies might also evaluate aspects of the visual production of each syllable to investigate what features might determine the degree of redundancy in the signal.

References

Andrews, B. (2007). Auditory and visual information facilitating speech integration. Undergraduate honors thesis, The Ohio State University.

Grant, K.W. (2002). Measures of auditory-visual integration for speech understanding: A theoretical perspective (L). *The Journal of the Acoustical Society of America*, 112 (1), 30-33.

Grant, K.W. and Seitz, P.F. (1998). Measures of auditory-visual integration in nonsense syllables and sentences. *The Journal of the Acoustical Society of America*, 104, 2438-2450.

Green, K.P. (1998). The use of auditory and visual information during phonetic processing: Implications for theories of speech perception. In Campbell, R., Dodd, B., and Burnham, D. (Eds.), *Hearing by eye II*: *Advances in the psychology of speechreading and auditory-visual speech*. East Sussex, UK: Psychology Press, Ltd.

Huffman, C. (2007). The role of auditory information in audiovisual speech integration. Undergraduate honors thesis, The Ohio State University.

Hungerford, M. (2007). The role of information redundancy in audiovisual speech integration. Undergraduate honors thesis, The Ohio State University.

Jackson, P.L. (1988). The theoretical minimal unit for visual speech perception: Visemes and coarticulation. *The Volta Review*, 90 (5), 99-114.

McGurk, H. and MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264, 746-748.

Remez, R.E., Rubin, P.E., Pisoni, D.B. and Carrell, T.D. (1981). Speech perception without traditional speech cues. *Science*, 212, 947-950.

Shannon, R.V., Zeng, F.G., Kamath, V., Wygonski, J., and Ekelid, M. (1995). Speech recognition with primarily temporal cues. *Science*, 270, 303-304.

Smith, Z.M., Oxenham, A.O., and Delgutte, B. (2002). Chimaeric sounds reveal dichotomies in auditory perception. *Nature*, 416, 87-90.

List of Figures

Figure 1: Behavioral Confusion Matrix for Talker JK in the 2-Channel Condition

Figure 2: Behavioral Confusion Matrix for Talker LG in the 2-Channel Condition

Figure 3: Behavioral Confusion Matrix for Talker JK in the 4-Channel Condition

Figure 4: Behavioral Confusion Matrix for Talker LG in the 4-Channel Condition

Figure 5: Formant Tracks for Talker JK producing the stimuli “cat” and “pat” in the 2-Channel Condition

Figure 6: Formant Tracks for Talker JK producing the stimuli “bat” and “gat” in the 2-Channel Condition

Figure 7: Formant Tracks for Talker JK producing the stimuli “cat”, “pat”, and “tat” in the 2-Channel Condition

Figure 8: Formant Tracks for Talker LG producing the stimuli “cat” and “gat” in the 2-Channel Condition

Figure 9: Formant Tracks for Talker LG producing the stimuli “cat”, “pat”, and “tat” in the 2-Channel Condition

Figure 10: Formant Tracks for Talker JK producing the stimuli “bat”, “gat”, and “pat” in the 4-Channel Condition

Figure 11: Formant Tracks for Talker JK producing the stimuli “pat” and “tat” in the 4-Channel Condition

Figure 12: Formant Tracks for Talker LG producing the stimuli “cat”, “pat”, and “tat” in the 4-Channel Condition

Figure 13: Percent Correct for Talker JK & LG in the 2-Channel Condition

Figure 14: Percent Correct for Talker JK & LG in the 4-Channel Condition

Figure 15: Behavioral Confusion Matrix for Talker JK in the 2-Channel Auditory + Visual Condition

Figure 16: Behavioral Confusion Matrix for Talker LG in the 2-Channel Auditory + Visual Condition

Figure 17: Behavioral Confusion Matrix for Talker JK in the 4-Channel Auditory + Visual Condition

Figure 18: Behavioral Confusion Matrix for Talker LG in the 4-Channel Auditory + Visual Condition

Behavioral Confusion Matrix for Talker JK in the 2-Channel Auditory-only Condition
Figure 1

		RESPONSES															
		BAT	**PAT**	**MAT**	**GAT**	**CAT**	**ZAT**	**TAT**	**SAT**	**AT**	**RAT**	**THAT**	**FAT**	**HAT**	**BGAT**	**NAT**	**DAT**
STIMULI	**BAT**	**0.5**	0.01	0.02	0.03				0.01	0.15		0.05	0.03	0.19	0.01		
	PAT	0.03	**0.37**		0.02	0.26		0.03			0.01	0.08	0.08	0.12			
	MAT			**0.73**		0.05						0.05		0.13		0.03	
	GAT	0.73			**0.12**	0.03				0.01		0.05	0.03	0.02			0.03
	CAT	0.01	0.14			**0.63**		0.11	0.01				0.08	0.03		0.01	
	ZAT	0.08	0.13	0.05	0.08	0.05	**0**		0.03	0.03	0.05	0.25	0.08	0.2			
	TAT	0.03	0.28		0.03	0.48		**0.05**				0.05	0.05	0.05			
	SAT	0.03	0.05					0.03	**0.15**			0.03	0.73				

Behavioral Confusion Matrix for Talker LG in the 2-Channel Auditory-only Condition
Figure 2

		RESPONSES																
		BAT	**PAT**	**MAT**	**GAT**	**CAT**	**ZAT**	**TAT**	**SAT**	**HAT**	**FAT**	**THAT**	**NAT**	**DAT**	**BRAT**	**BLAT**	**THAET**	**AT**
STIMULI	**BAT**	**0.67**	0.01	0.1	0.09	0.01		0.01	0.01		0.07	0.03						
	PAT		**0.86**	0.01		0.09				0.02	0.02	0.01						
	MAT		0.03	**0.78**		0.03				0.05			0.13					
	GAT	0.1	0.04	0.01	**0.57**	0.18		0.02			0.01			0.03	0.03	0.03		
	CAT	0.01	0.11		0.02	**0.8**		0.05		0.01								
	ZAT	0.25	0.03	0.15	0.03		**0.13**			0.08	0.18	0.18						
	TAT	0.03	0.23		0.05	0.45		**0.25**										
	SAT	0.05	0.05						**0.15**		0.65	0.03					0.05	0.03

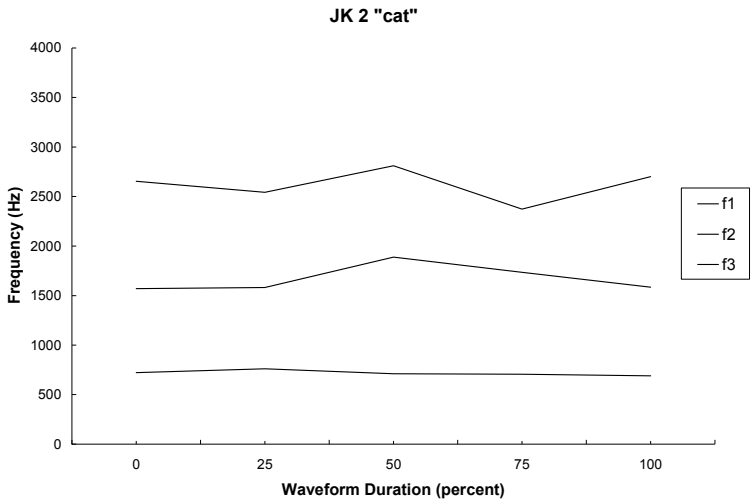
Behavioral Confusion Matrix for Talker JK in the 4-Channel Auditory-only Condition
Figure 3

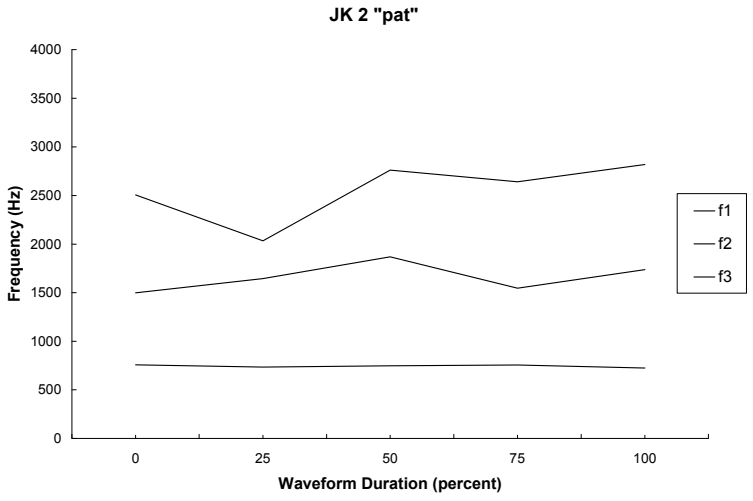
		RESPONSES															
		BAT	**PAT**	**MAT**	**GAT**	**CAT**	**ZAT**	**TAT**	**SAT**	**HAT**	**FAT**	**DAT**	**THAT**	**NAT**	**AT**	**SHAT**	**THAET**
STIMULI	**BAT**	**0.43**	0.07	0.08						0.25	0.04		0.05	0.01	0.07		
	PAT	0.01	**0.71**	0.01		0.02				0.26							
	MAT	0.05	0.08	**0.73**			0.05			0.08	0.03						
	GAT	0.21	0.16	0.02	**0.18**	0.03		0.04		0.02		0.29	0.03	0.02			
	CAT		0.09	0.01		**0.67**		0.1		0.11	0.01		0.01				
	ZAT	0.18		0.03			**0.1**		0.05	0.08	0.08	0.03	0.45				0.03
	TAT		0.49			0.12		**0.32**		0.05	0.02						
	SAT	0.05		0.03			0.03		**0.13**		0.63		0.08			0.05	0.03

Behavioral Confusion Matrix for Talker LG in the 4-Channel Auditory-only Condition
Figure 4

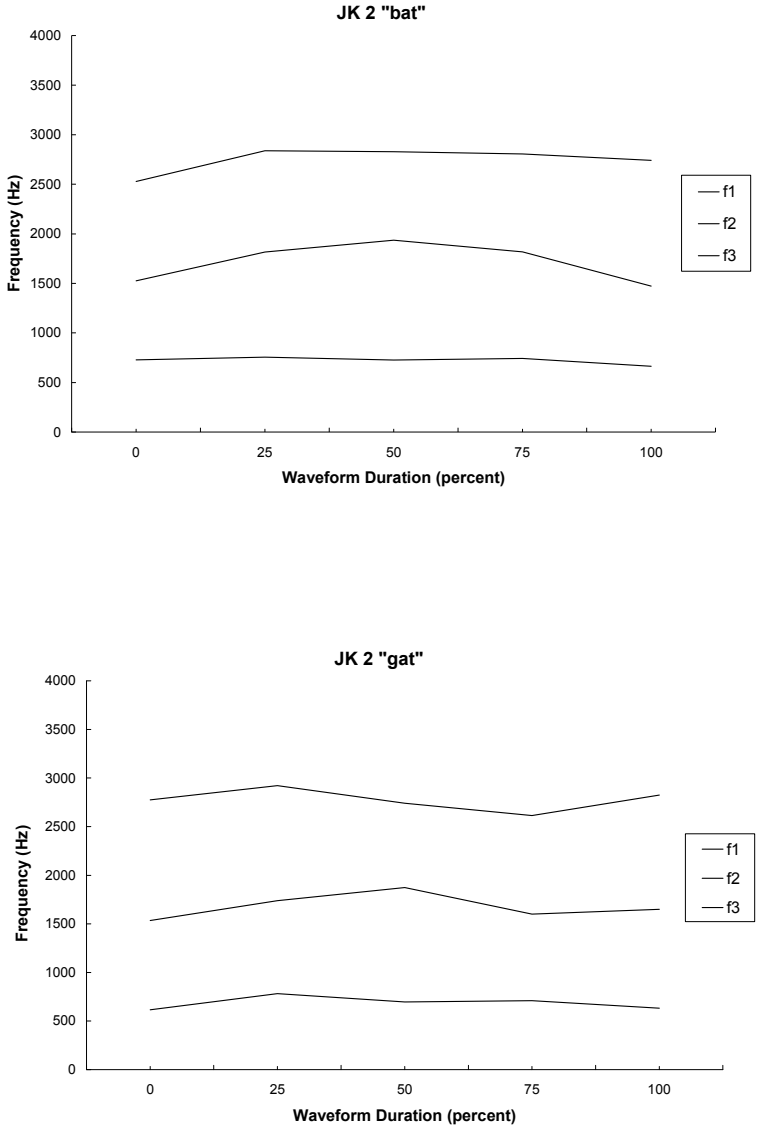
		RESPONSES															
		BAT	**PAT**	**MAT**	**GAT**	**CAT**	**ZAT**	**TAT**	**SAT**	**HAT**	**NAT**	**FAT**	**THAT**	**DWAT**	**DAT**	**THAET**	**FLAT**
STIMULI	**BAT**	**0.56**		0.07			0.04			0.06		0.15	0.12				
	PAT		**0.92**			0.06				0.03							
	MAT			**0.88**		0.05		0.03	0.03		0.03						
	GAT	0.01			**0.66**	0.27		0.01					0.03	0.03	0.01		
	CAT	0.01	0.21	0.01	0.04	**0.65**		0.03				0.01	0.01	0.04			
	ZAT	0.13	0.03			0.03	**0.2**	0.01				0.08	0.53				
	TAT		0.13			0.13		**0.7**	0.05								
	SAT					0.03		0.03	**0.28**			0.59	0.03			0.03	0.03

Formant Tracks for Talker JK producing the stimuli “cat” and “pat” in the 2-Channel Condition
Figure 5



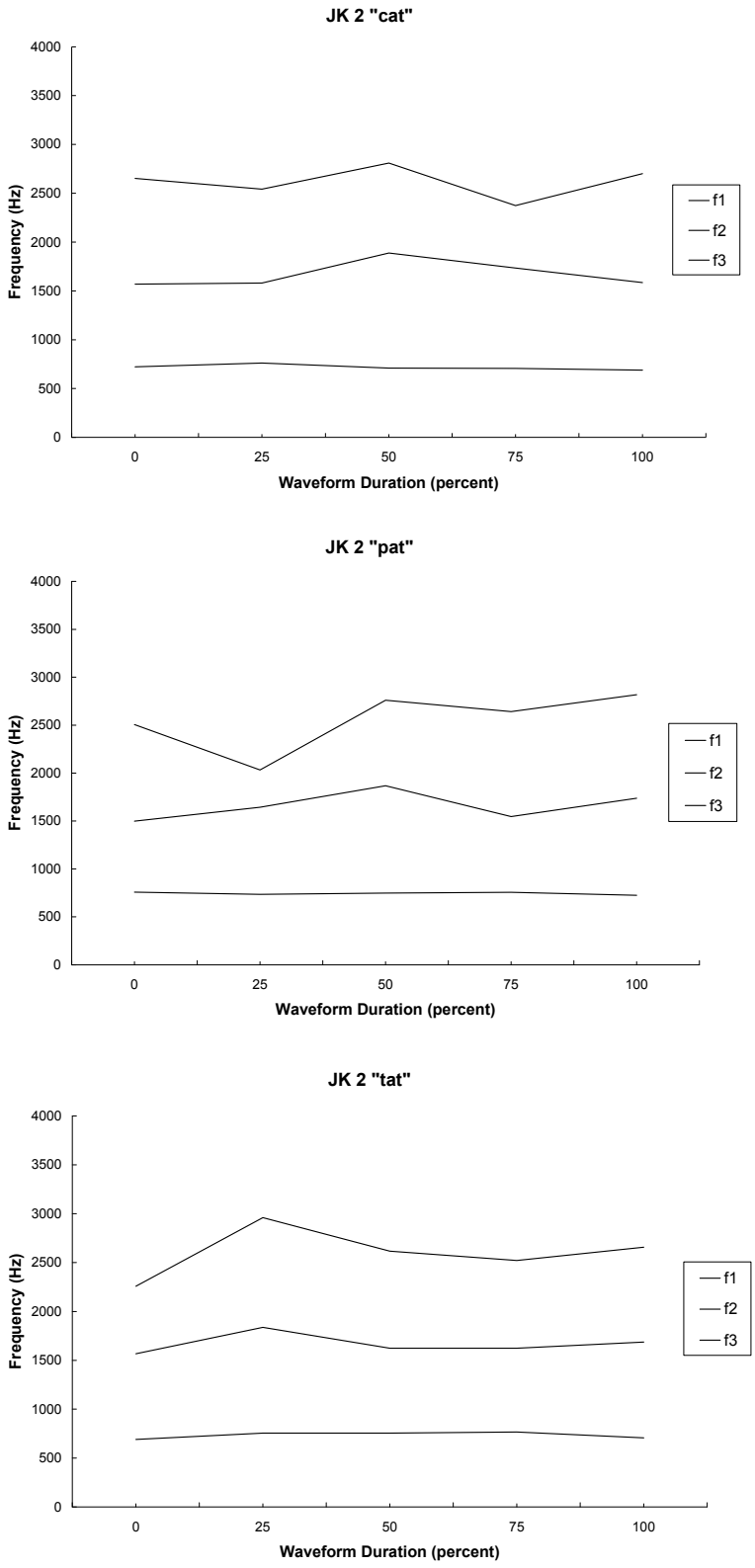


Formant Tracks for Talker JK producing the stimuli "bat" and "gat" in the 2-Channel Condition
Figure 6

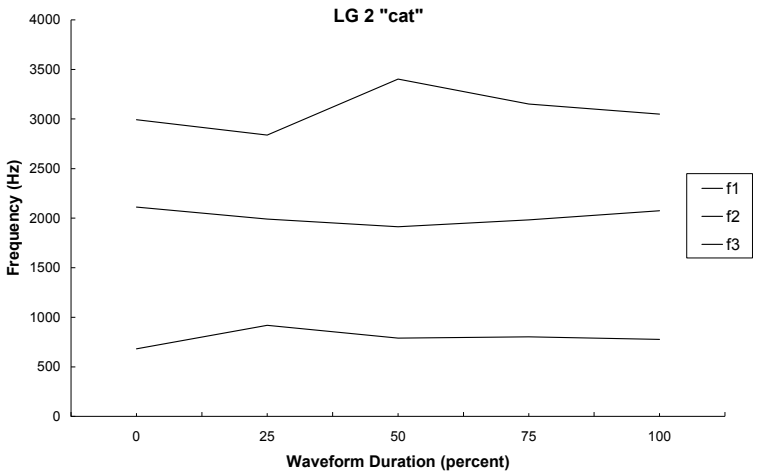


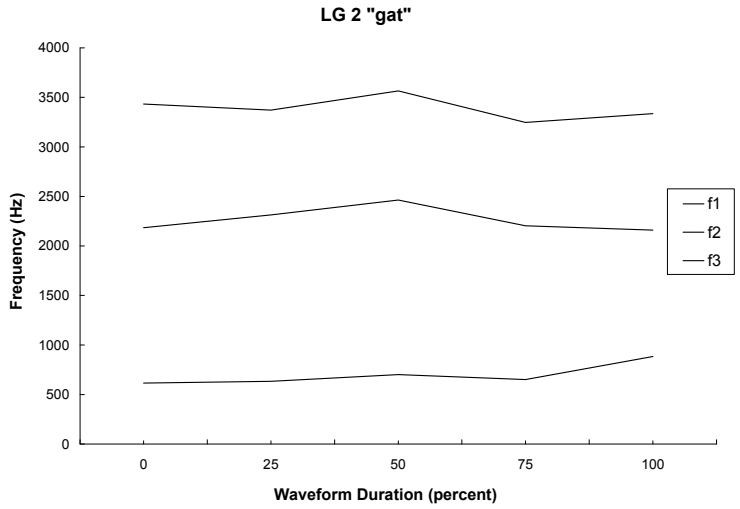
Formant Tracks for Talker JK producing the stimuli “cat,” “pat,” and “tat” in the 2-Channel Condition

Figure 7

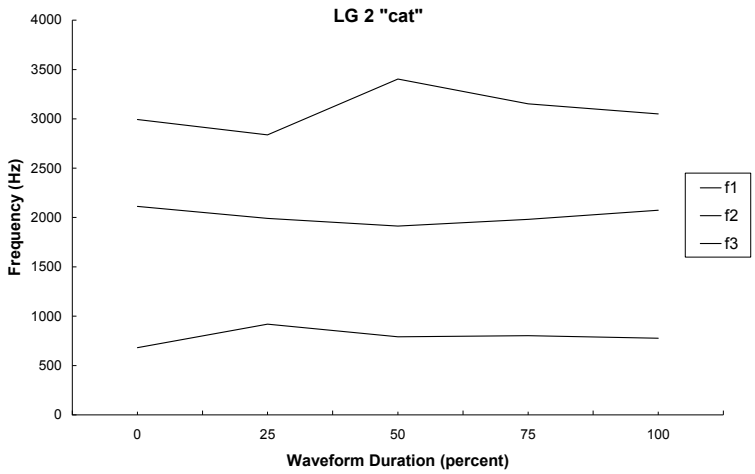


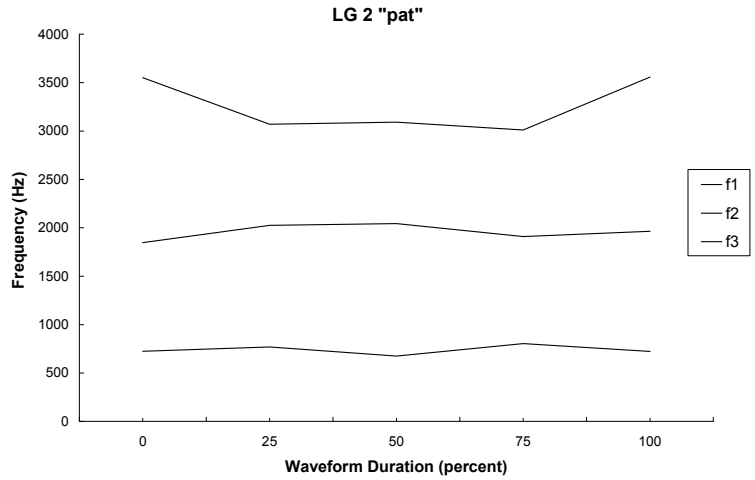
Formant Tracks for Talker LG producing the stimuli “cat” and “gat” in the 2-Channel Condition
Figure 8

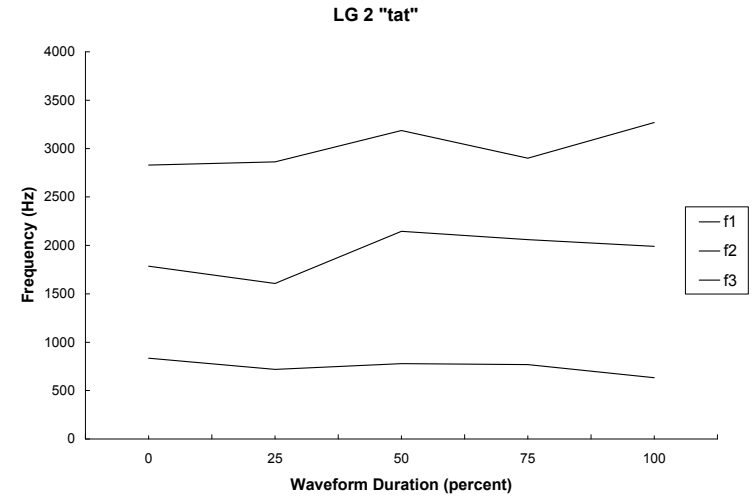




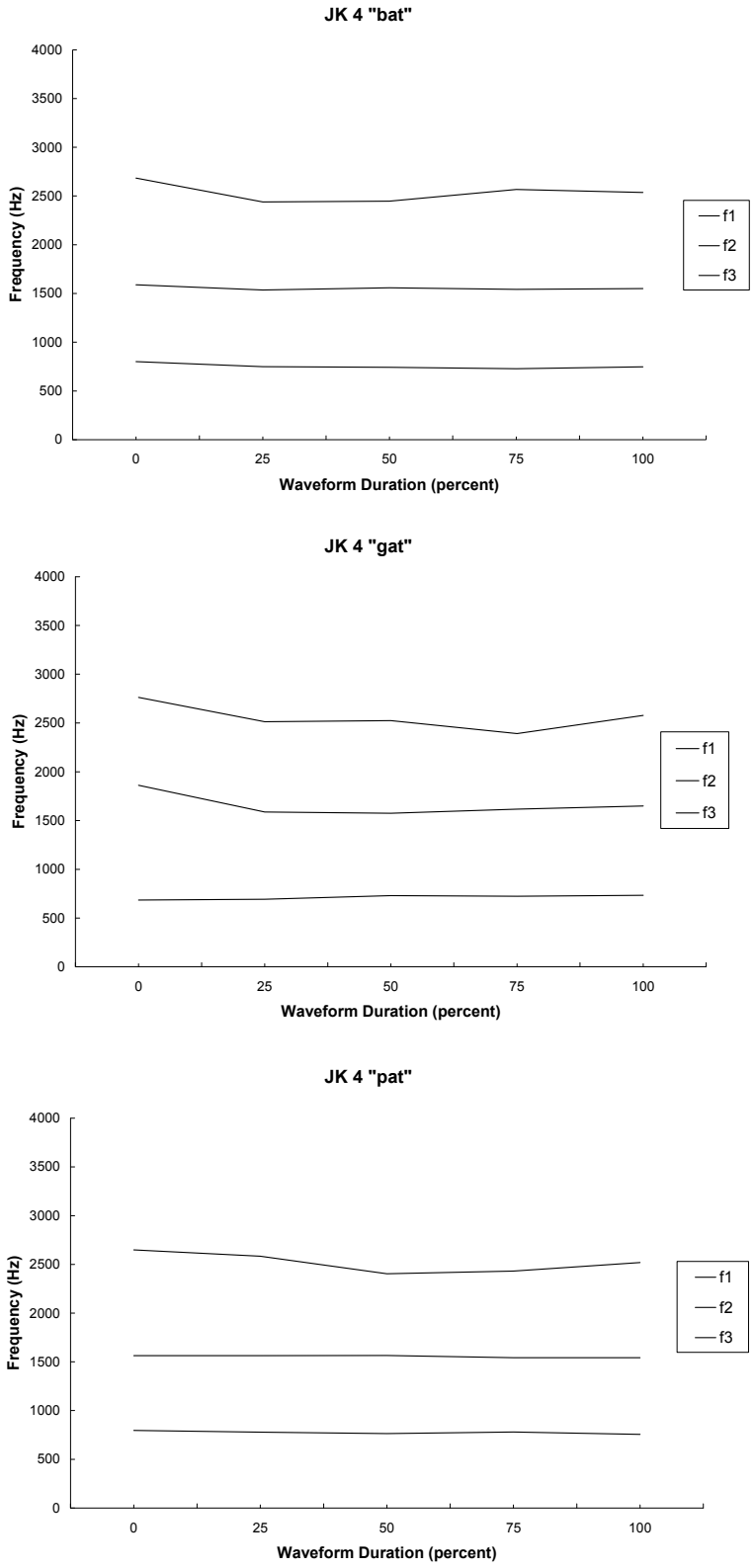
Formant Tracks for Talker LG producing the stimuli “cat,” “pat,” and “tat” in the 2-Channel Condition
Figure 9



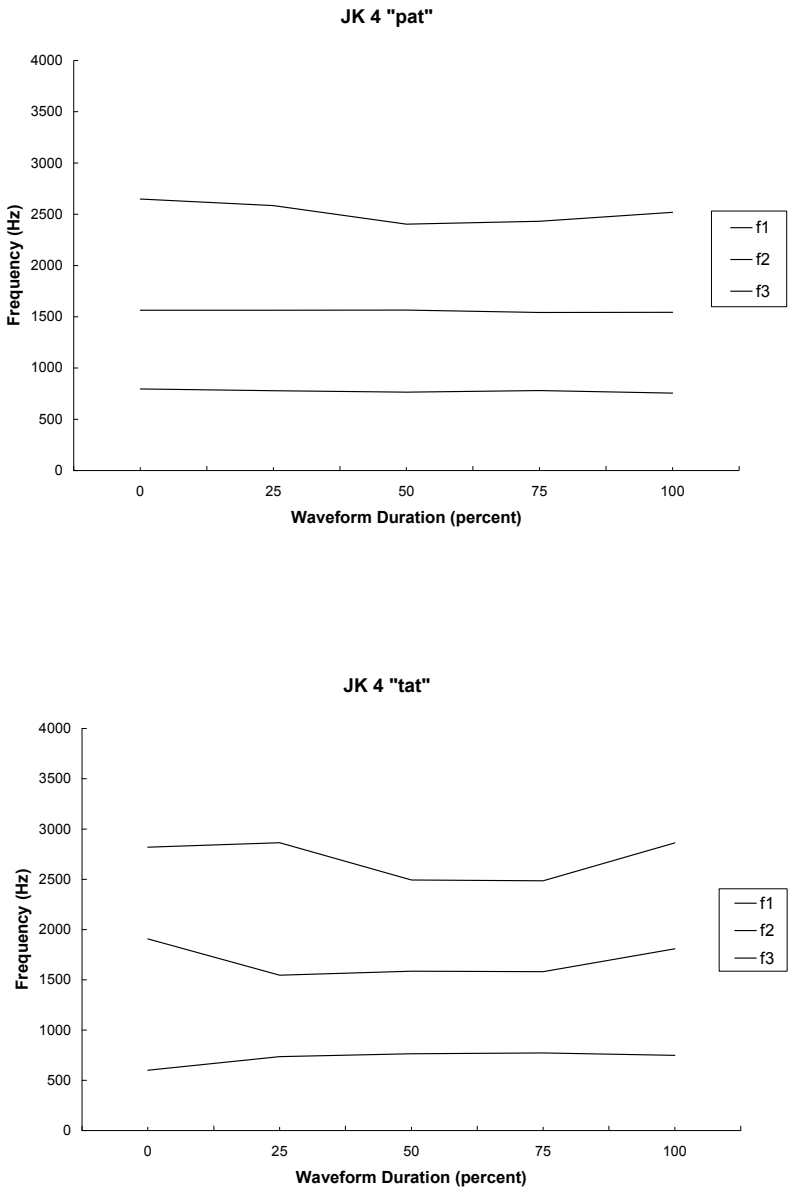




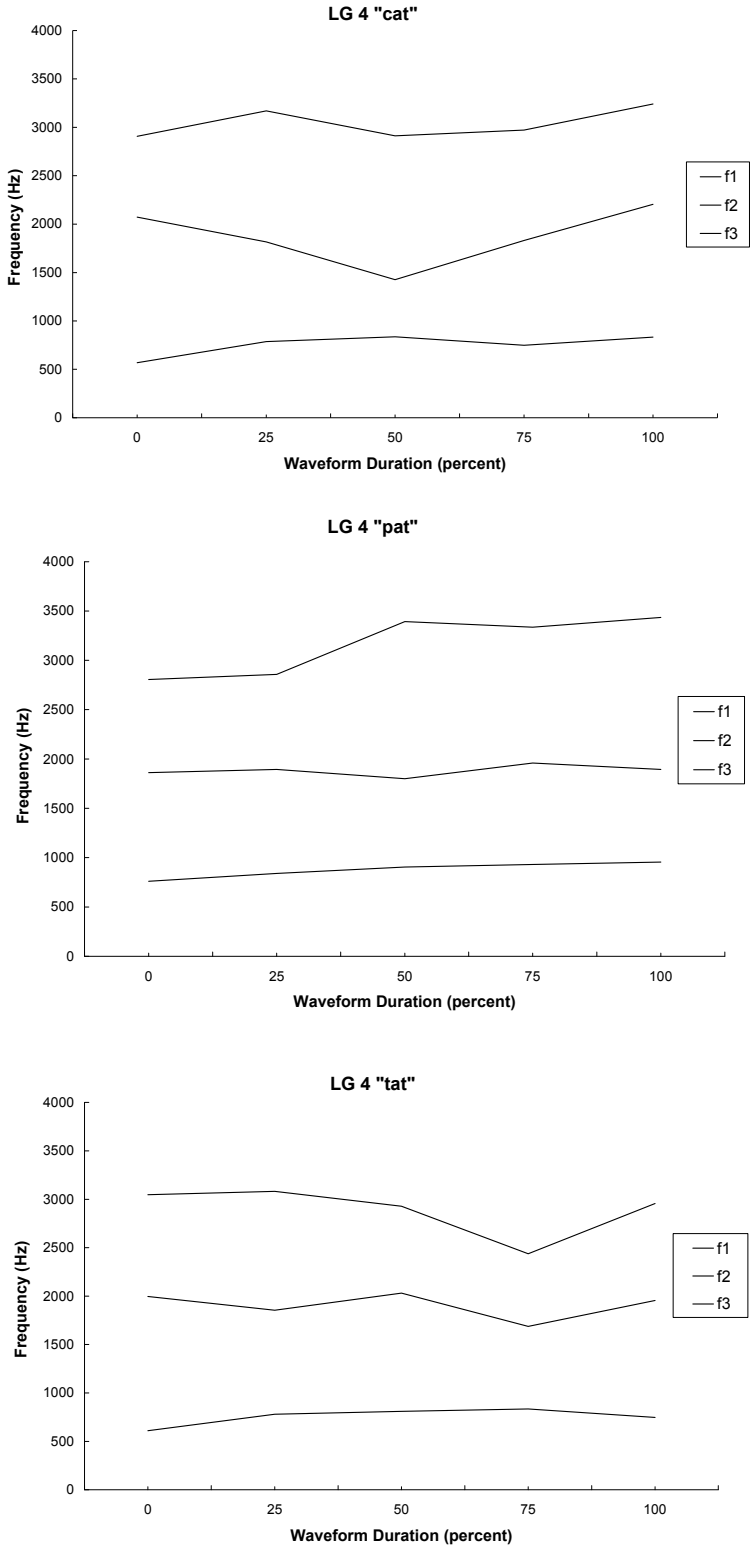
Formant Tracks for Talker JK producing the stimuli “bat,” “gat,” and “pat” in the 4-Channel Condition
Figure 10



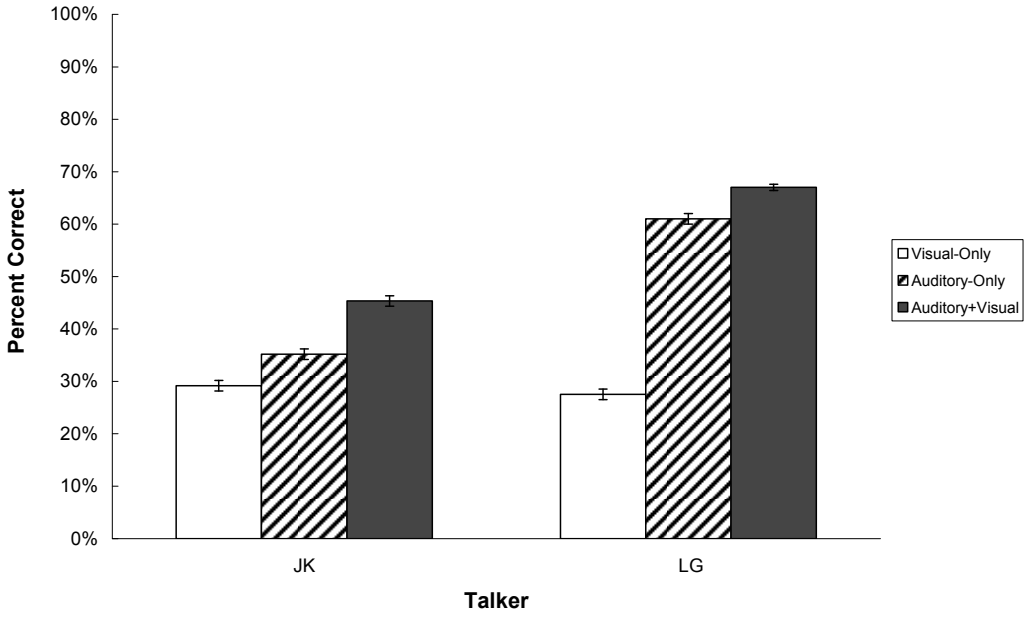
Formant Tracks for Talker JK producing the stimuli “pat” and “tat” in the 4-Channel Condition
Figure 11



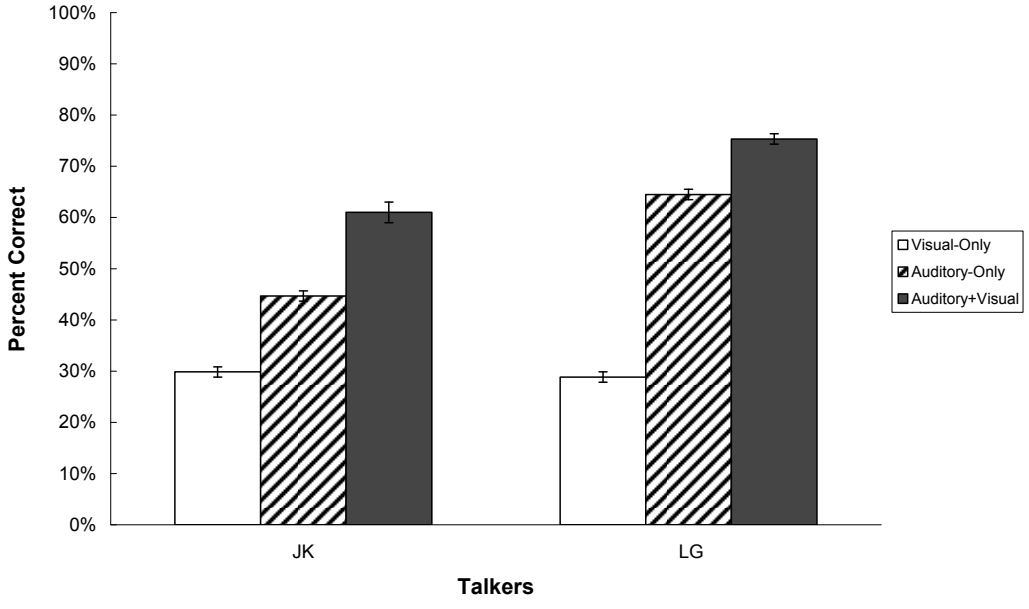
Formant Tracks for Talker LG producing the stimuli “cat,” “pat,” and “tat” in the 4-Channel Condition
Figure 12



Percent Correct for Talker JK & LG in the 2-Channel Condition
Figure 13



Percent Correct for Talker JK & LG in the 4-Channel Condition
Figure 14



Behavioral Confusion Matrix for Talker JK in the 2-Channel Auditory + Visual Condition
Figure 15

		RESPONSES														
		BAT	**PAT**	**MAT**	**GAT**	**CAT**	**ZAT**	**TAT**	**SAT**	**AT**	**FLAT**	**THAT**	**FAT**	**HAT**	**NAT**	**DAT**
	BAT	**0.6**	0.2	0.2									0			
	PAT	0.07	**0.86**											0.07		
	MAT	0.15		**0.82**											0.03	
STIMULI	**GAT**	0.05			**0.35**	0.08				0.1		0.18		0.2		0.05
	CAT		0.08			**0.78**		0.13					0.03			
	ZAT	0.05	0.05	0.03	0.03	0.08	**0**		0.04	0.03		0.13		0.43	0.05	
	TAT					0.85		**0.08**				0.05	0.03			
	SAT	0.05				0.02			**0.33**		0.05	0.08	0.5			

Behavioral Confusion Matrix for Talker LG in the 2-Channel Auditory + Visual Condition
Figure 16

		RESPONSES															
		BAT	**PAT**	**MAT**	**GAT**	**CAT**	**ZAT**	**TAT**	**SAT**	**AT**	**RAT**	**THAT**	**FAT**	**HAT**	**TWAT**	**NAT**	**DAT**
	BAT	**0.81**		0.06						0.03	0.06			0.03			
	PAT	0.03	**0.89**			0.03					0.03				0.03		
	MAT	0.03		**0.95**												0.03	
STIMULI	**GAT**				**0.55**	0.45											
	CAT					**0.94**		0.03									0.03
	ZAT	0.03			0.13	0.05	**0.2**	0.05	0.08			0.35	0.03	0.03		0.03	0.05
	TAT		0.03			0.49	0.03	**0.44**					0.03				
	SAT					0.03	0.03	0.03	**0.67**			0.08	0.18				

Behavioral Confusion Matrix for Talker JK in the 4-Channel Auditory + Visual Condition
Figure 17

RESPONSES

STIMULI	BAT	PAT	MAT	GAT	CAT	ZAT	TAT	SAT	THAT	FAT	HAT	DAT
BAT	**0.83**	0.07	0.03								0.07	
PAT		**0.95**								0.05		
MAT	0.15	0.08	**0.75**						0.03			
GAT	0.03	0.03		**0.53**	0.15						0.08	0.2
CAT		0.13			**0.53**		0.03				0.3	
ZAT	0.05	0.03	0.03			**0.38**	0.03	0.03	0.35	0.05	0.05	0.03
TAT		0.08			0.15		**0.62**				0.13	0.03
SAT		0.03			0.03	0.03	0.03	**0.33**	0.1	0.45	0.03	

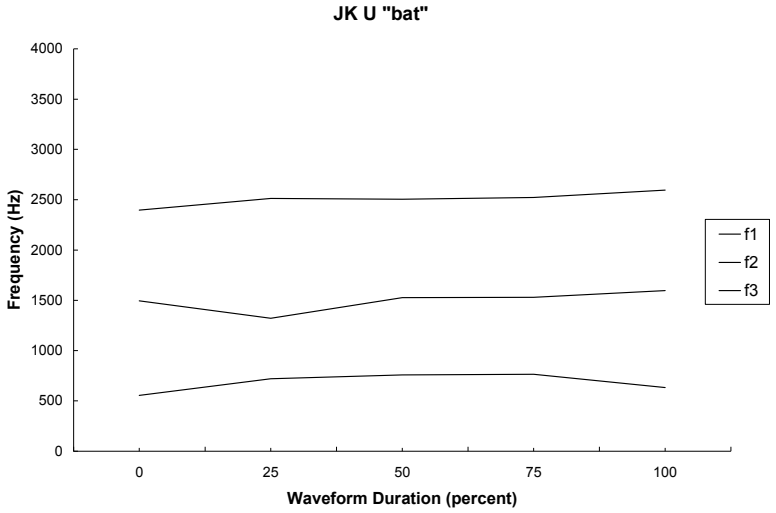
Behavioral Confusion Matrix for Talker LG in the 4-Channel Auditory + Visual Condition
Figure 18

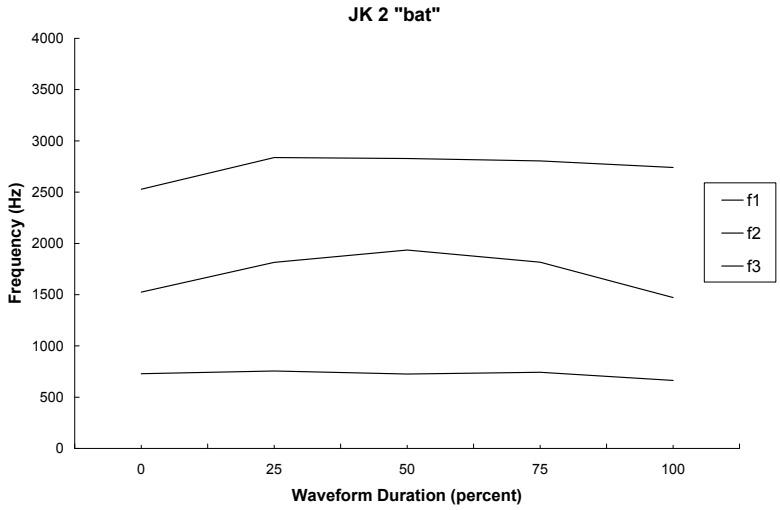
RESPONSES

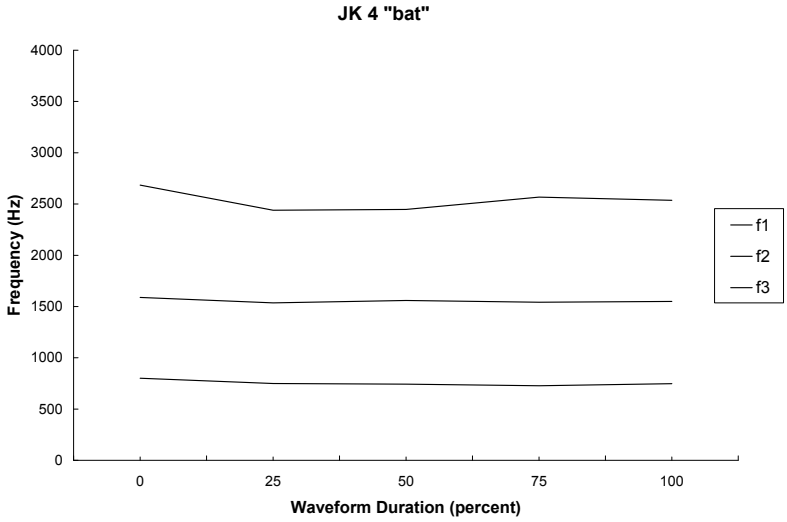
STIMULI	BAT	PAT	MAT	GAT	CAT	ZAT	TAT	SAT	AT	RAT	THAT	FAT	HAT	NAT	DAT
BAT	**0.81**	0.1	0.03							0.03		0.03			
PAT		**0.97**											0.03		
MAT		0.03	**0.93**								0.03			0.03	
GAT				**0.56**	0.44										
CAT		0.06		0.03	**0.83**		0.06					0.03			
ZAT	0.05		0.03			**0.44**		0.1	0.03		0.31	0.03			0.03
TAT		0.05			0.2		**0.75**								
SAT								**0.77**	0.03		0.05	0.15			

Appendix

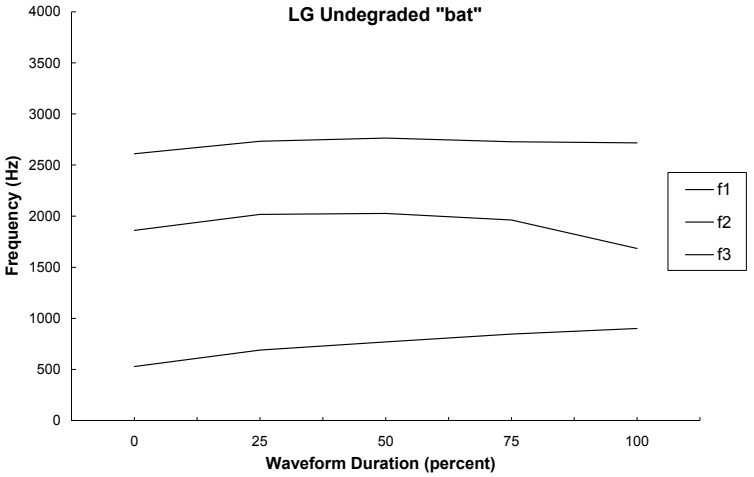
Formant Tracks for Talker JK producing the stimulus “bat”

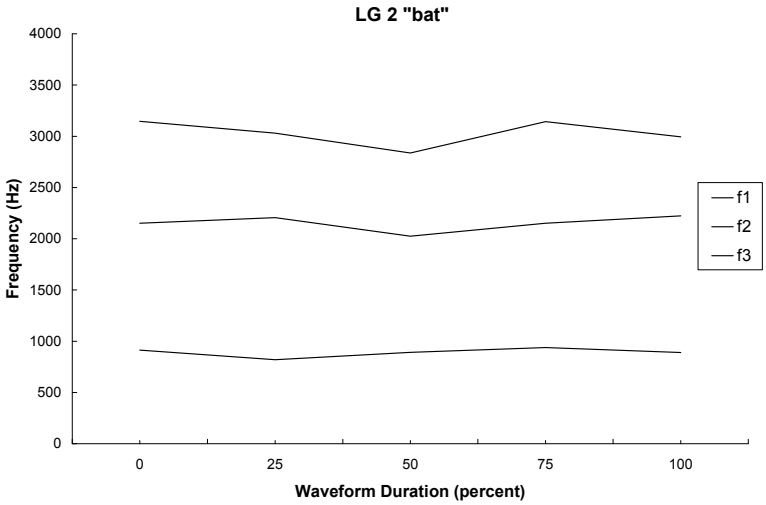


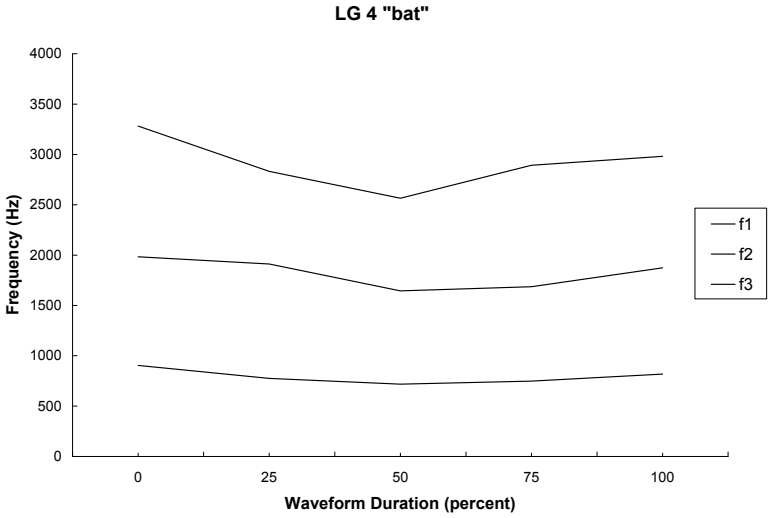




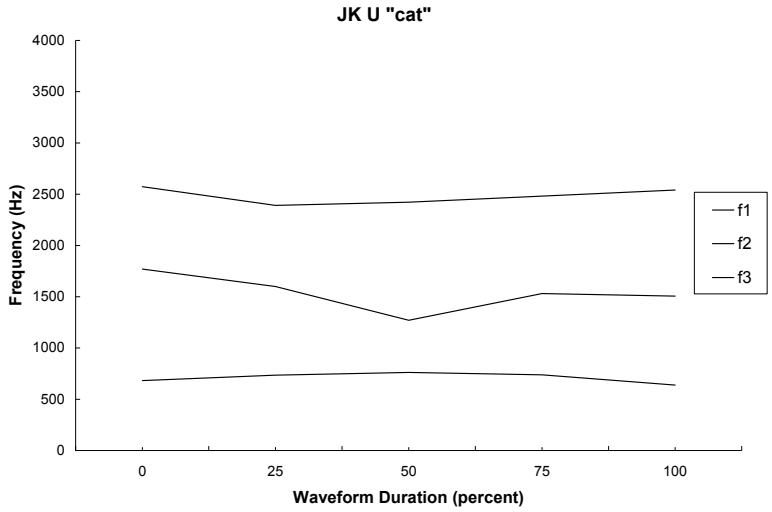
Formant Tracks for Talker LG producing the stimulus “bat”

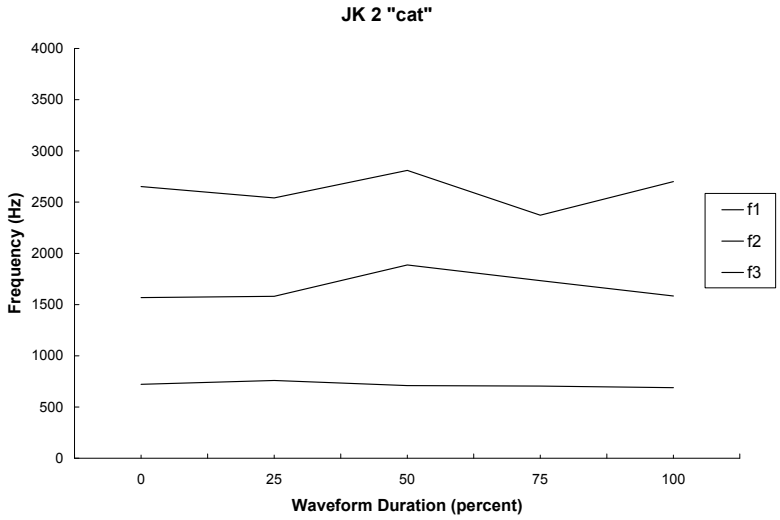


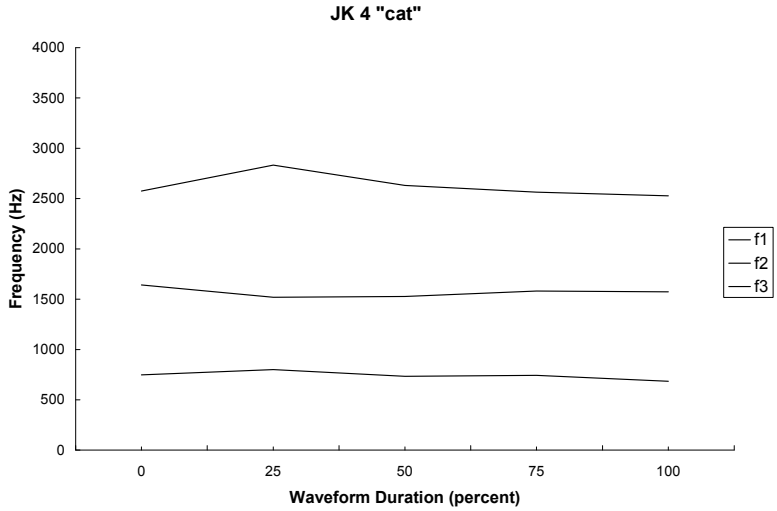




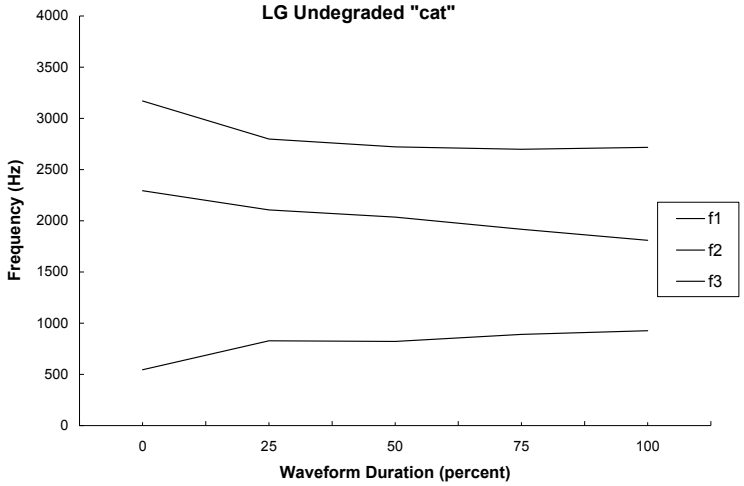
Formant Tracks for Talker JK producing the stimulus “cat”

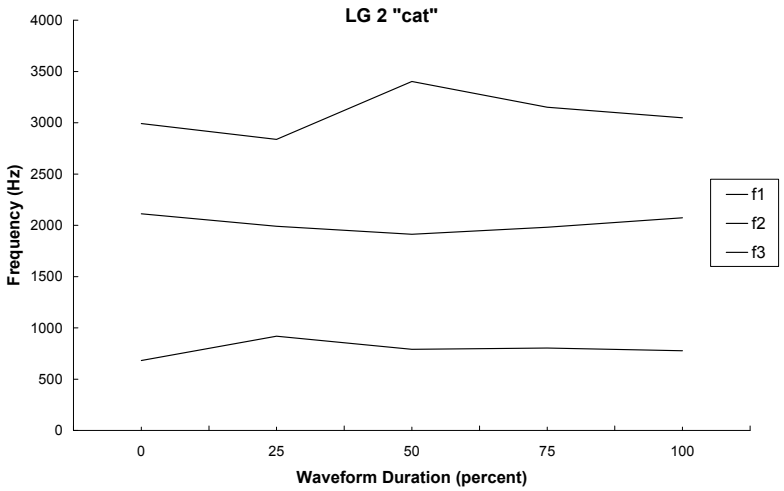


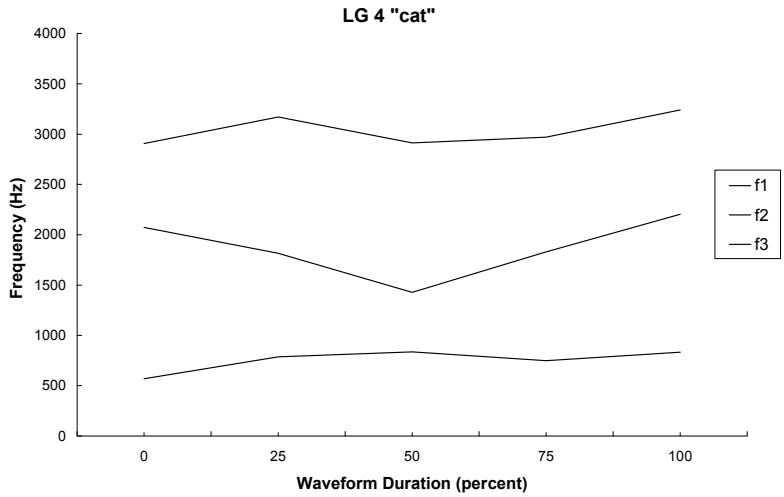




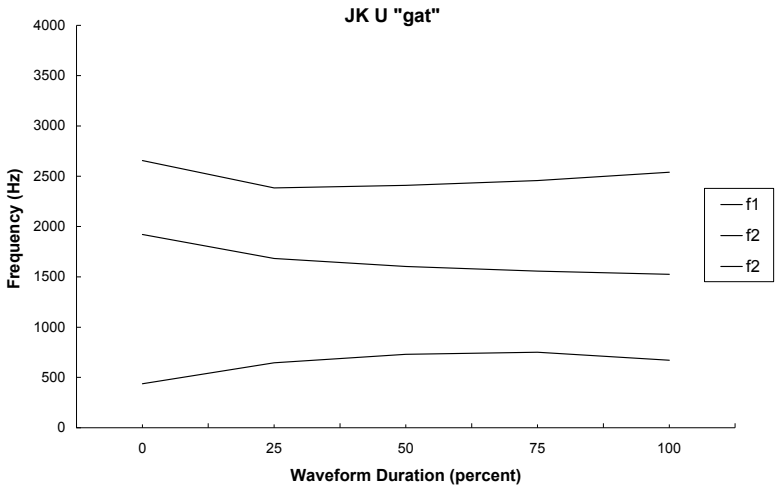
Formant Tracks for Talker LG producing the stimulus “cat”

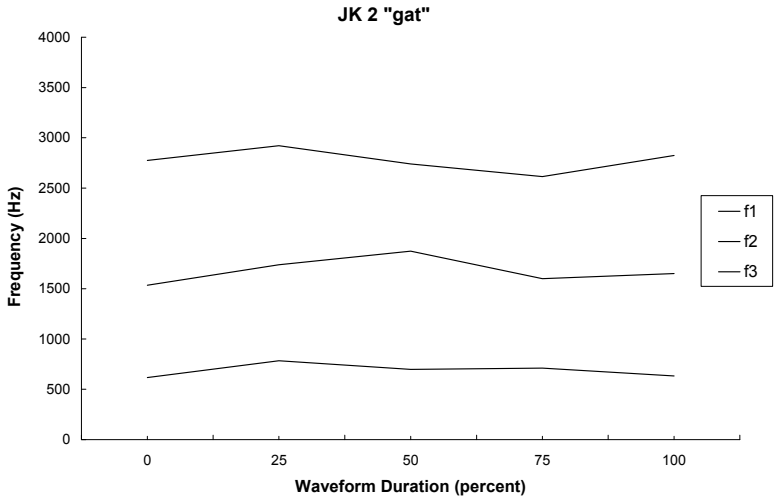


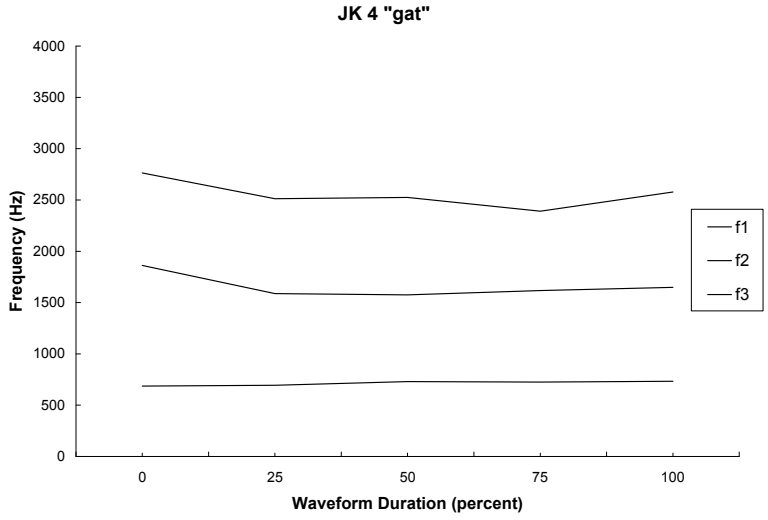




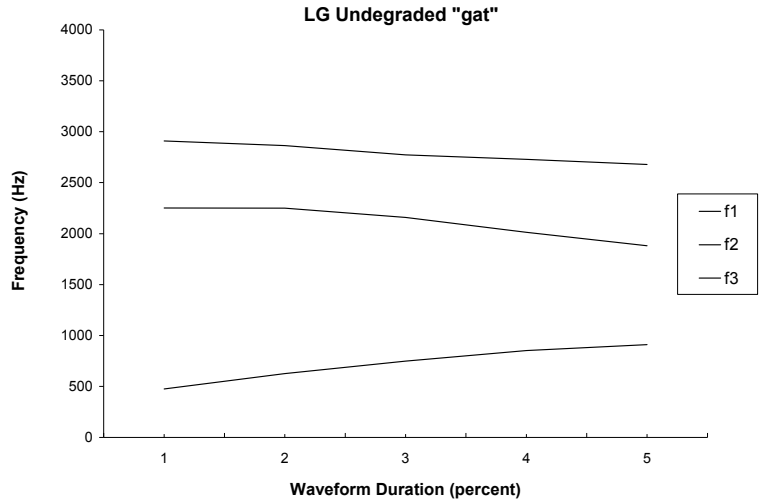
Formant Tracks for Talker JK producing the stimulus “gat”

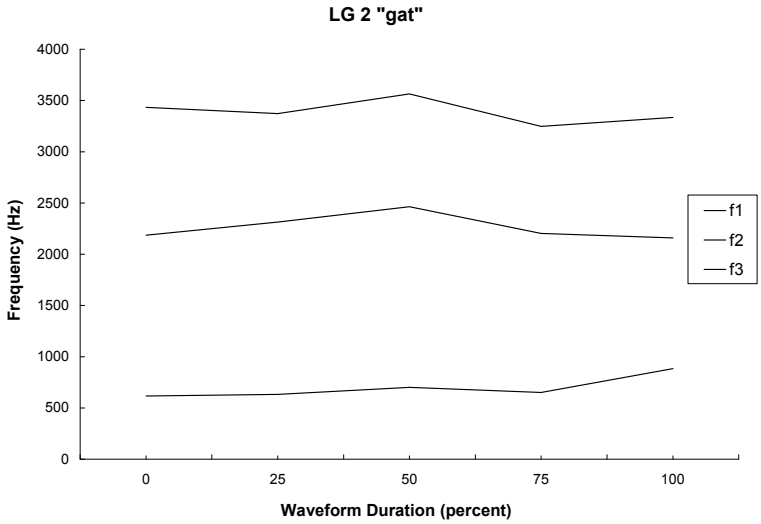


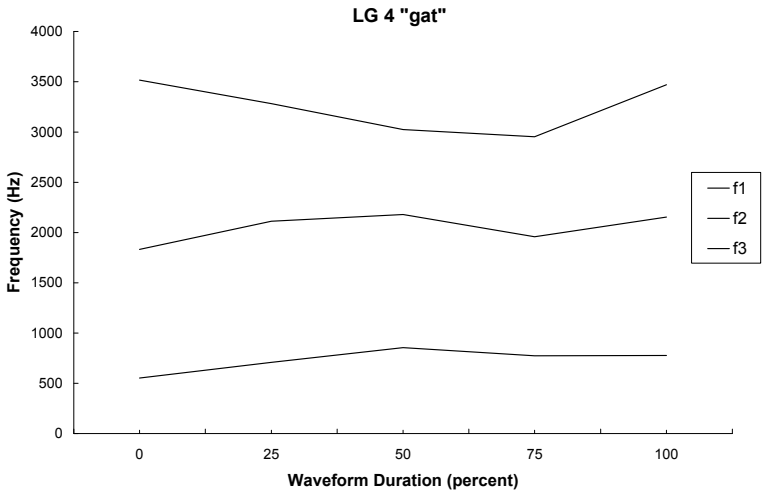




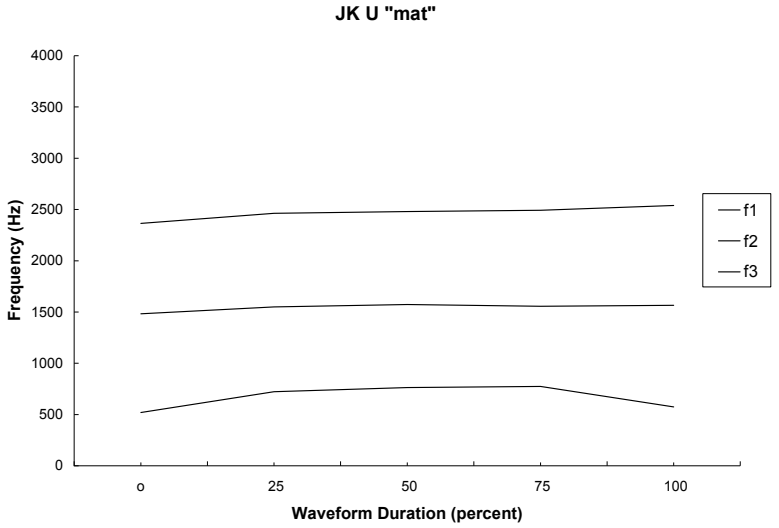
Formant Tracks for Talker LG producing the stimulus “gat”

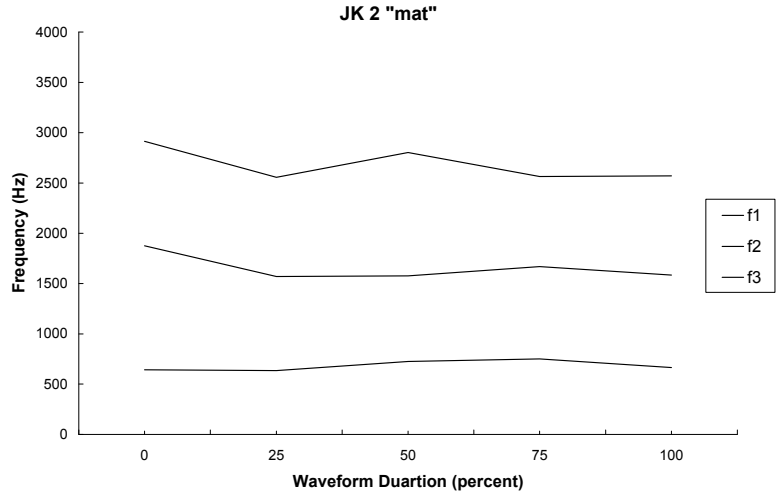


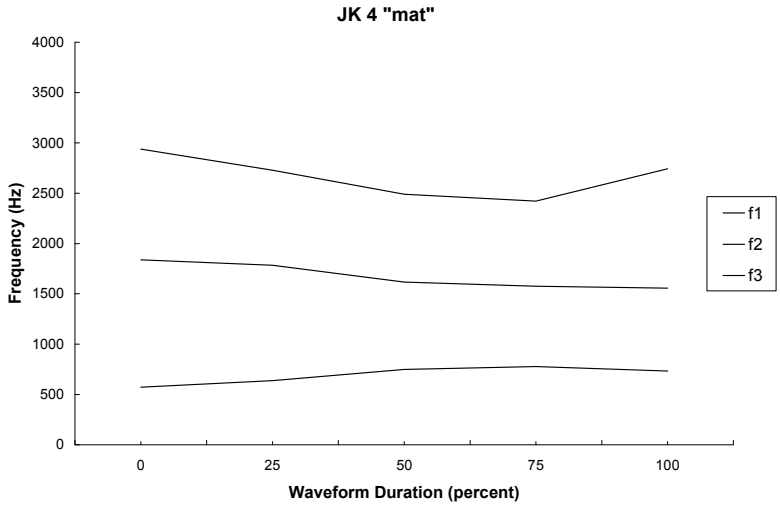




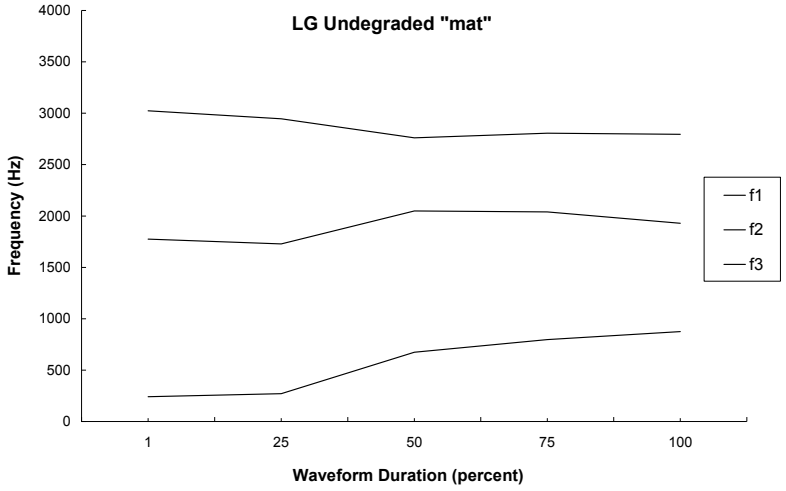
Formant Tracks for Talker JK producing the stimulus “mat”

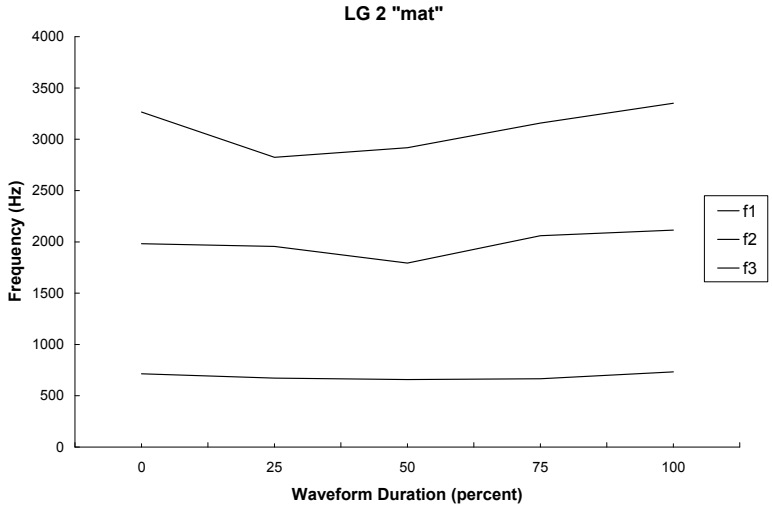


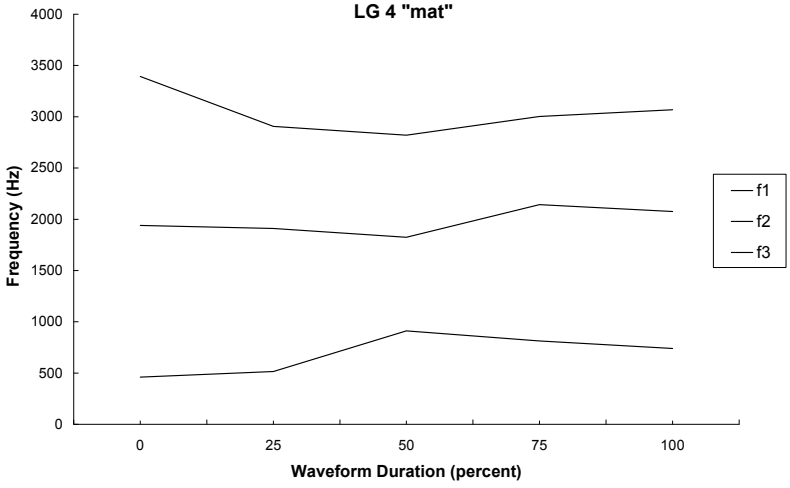




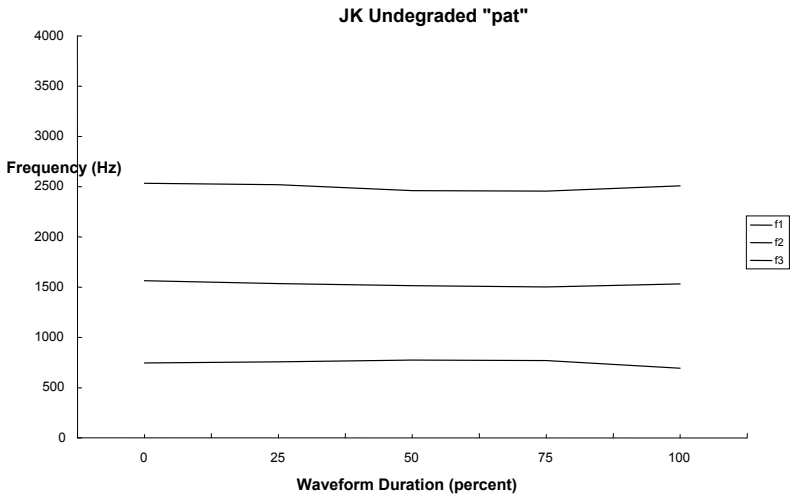
Formant Tracks for Talker LG producing the stimulus “mat”

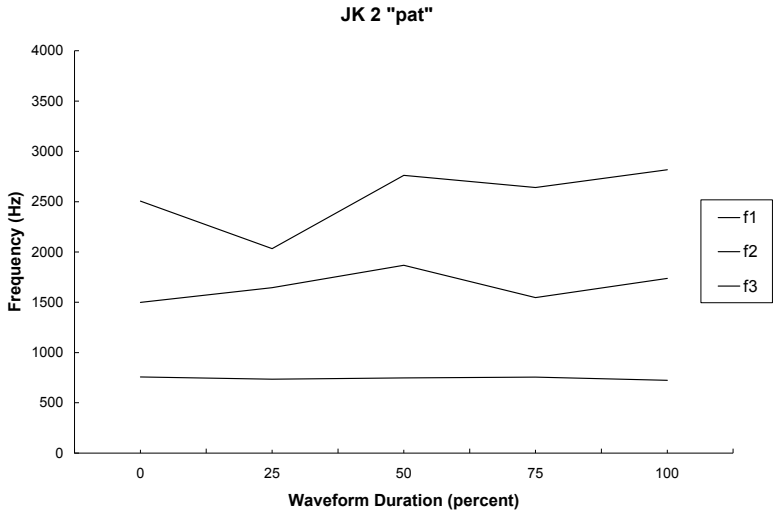


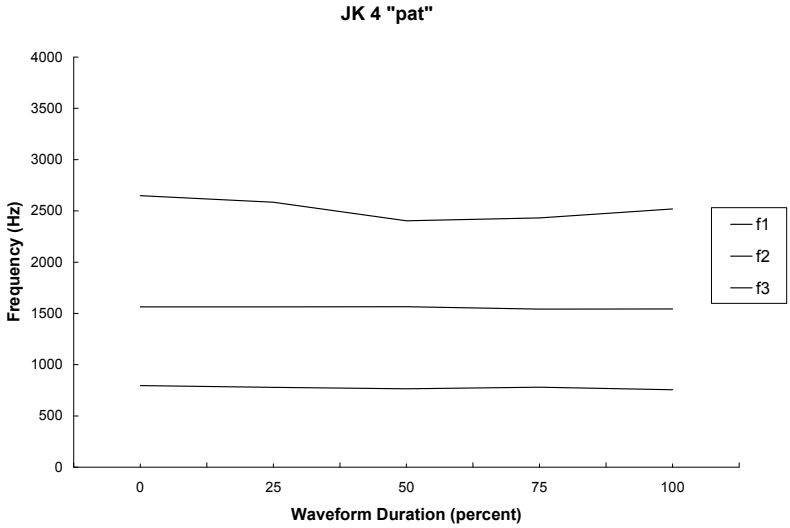




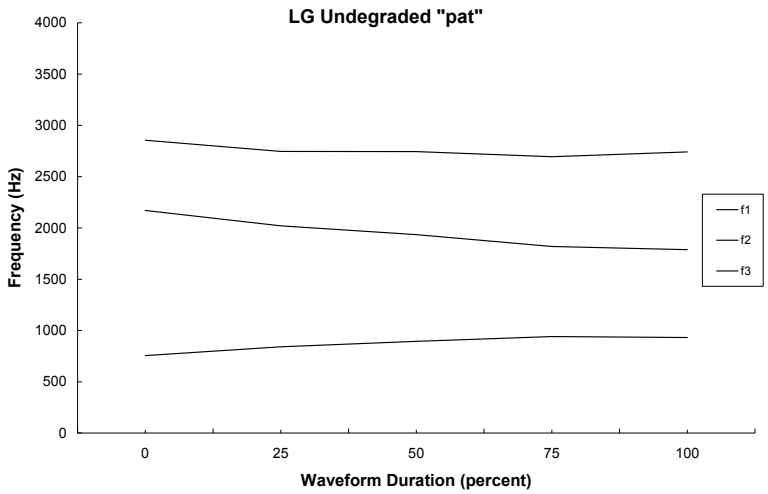
Formant Tracks for Talker JK producing the stimulus “pat”

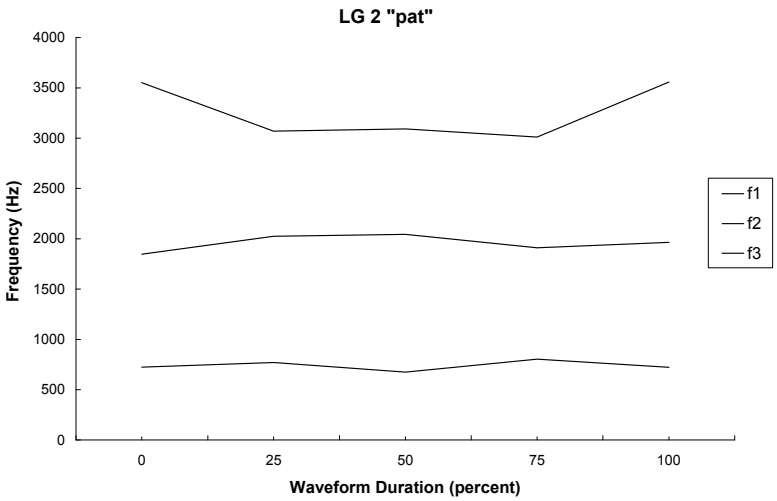


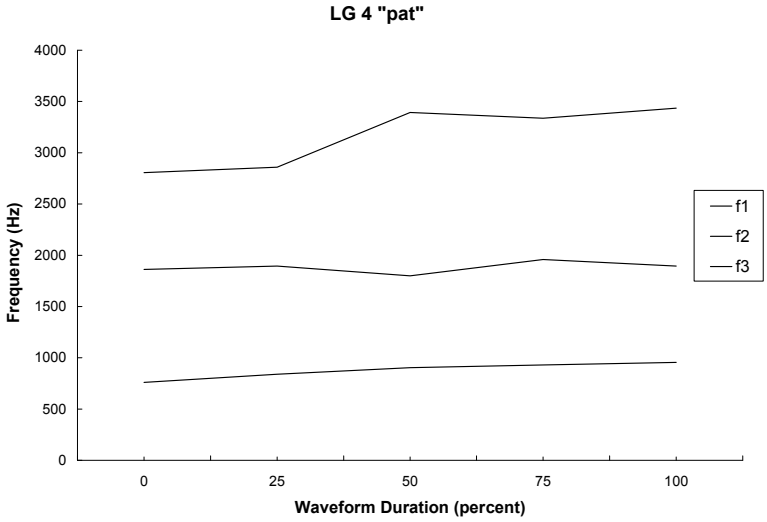




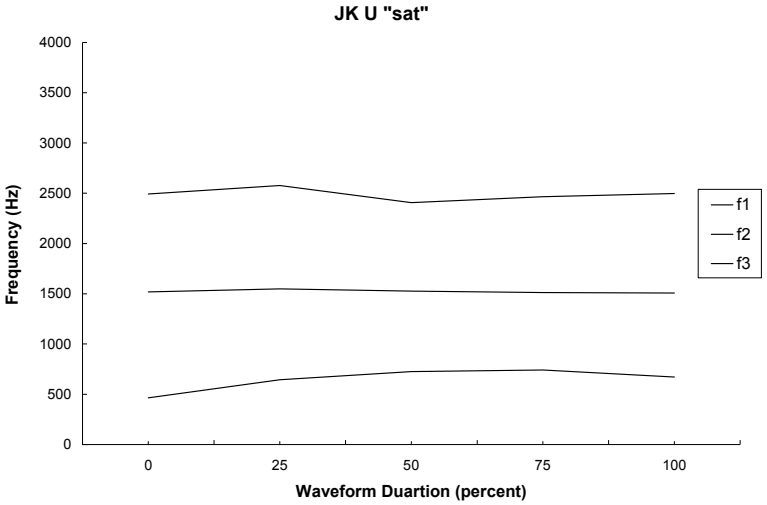
Formant Tracks for Talker LG producing the stimulus “pat”

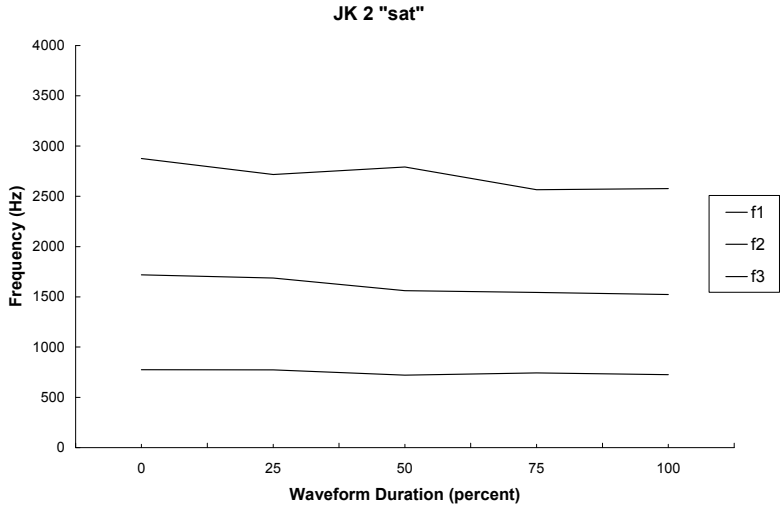


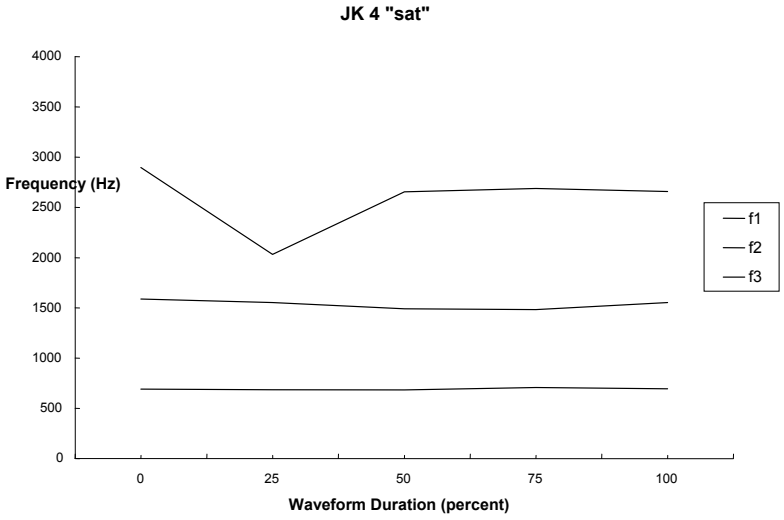




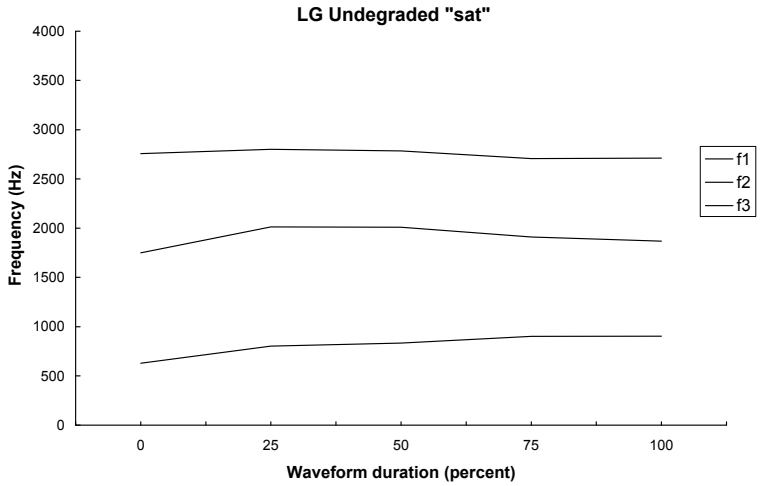
Formant Tracks for Talker JK producing the stimulus “sat”

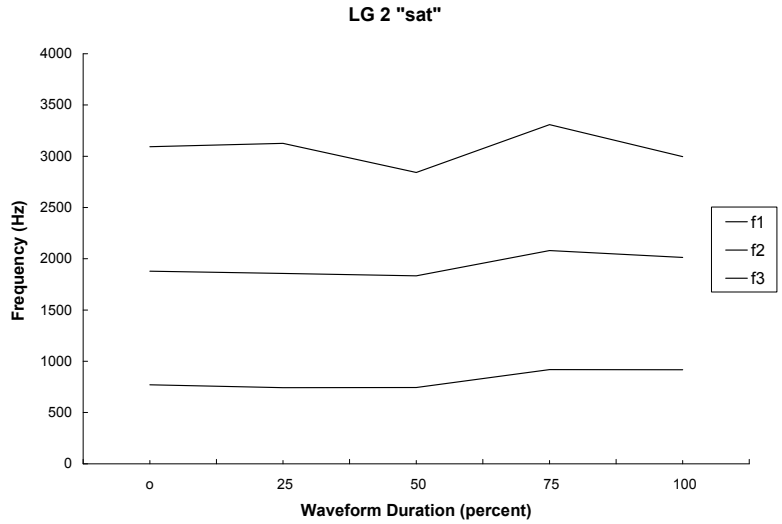


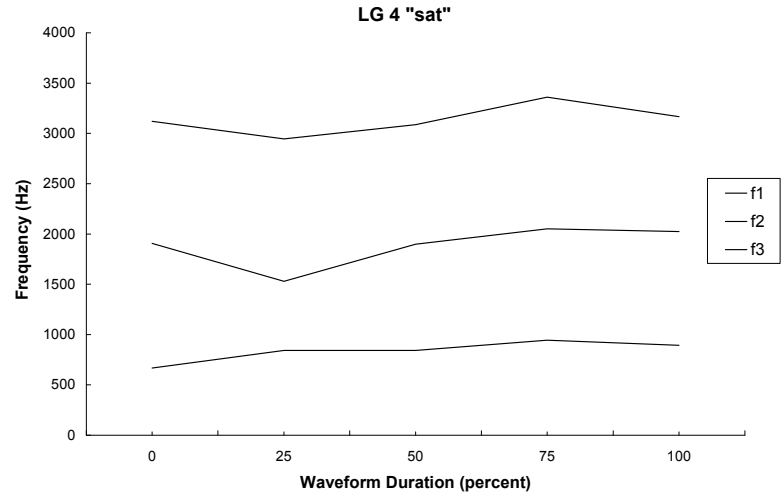




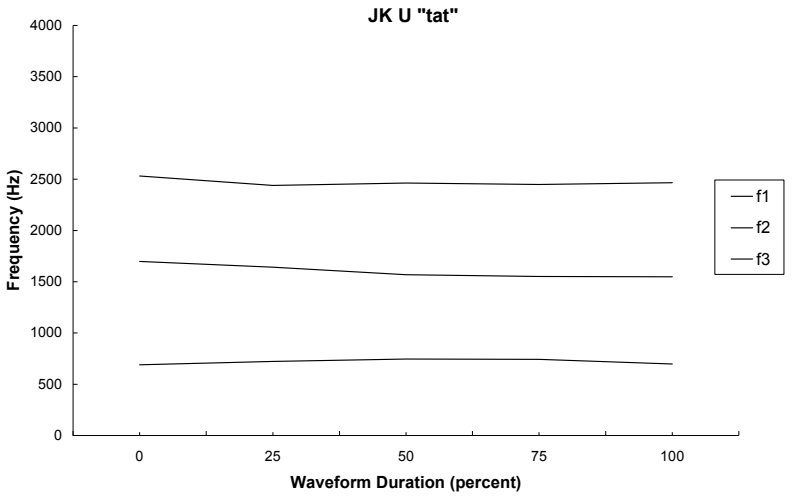
Formant Tracks for Talker LG producing the stimulus “sat”

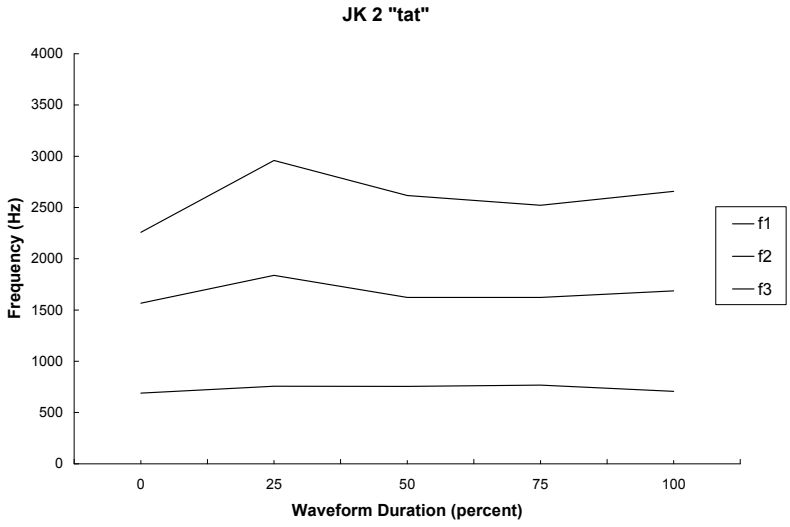


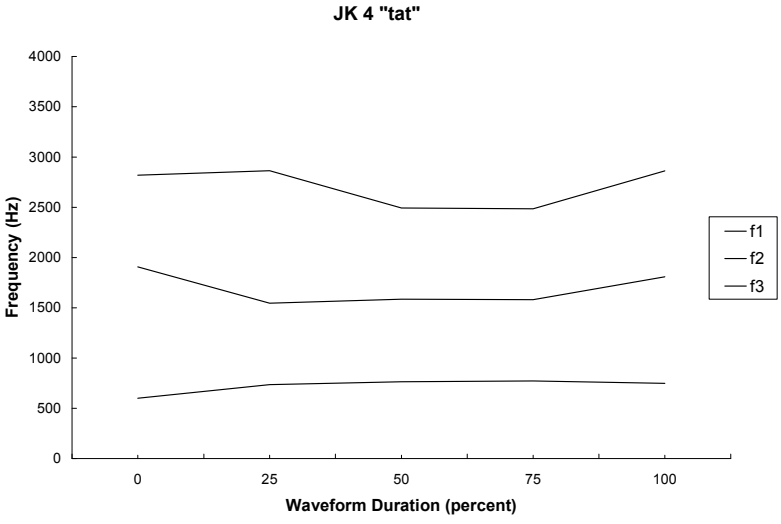




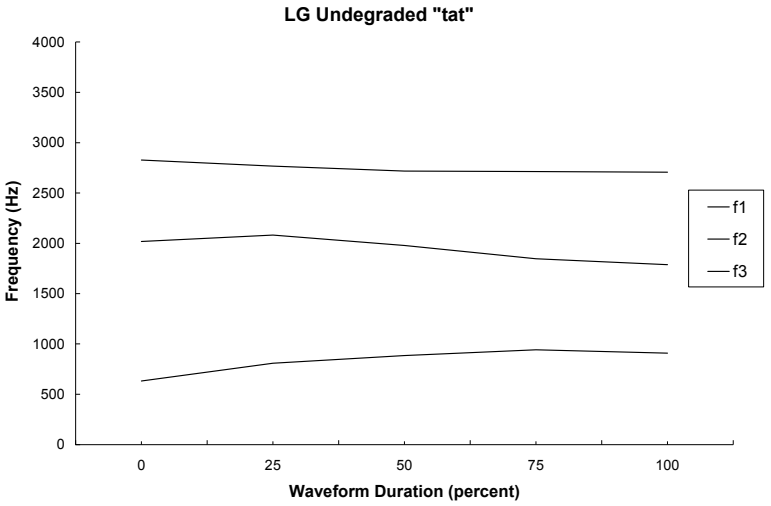
Formant Tracks for Talker JK producing the stimulus “tat”

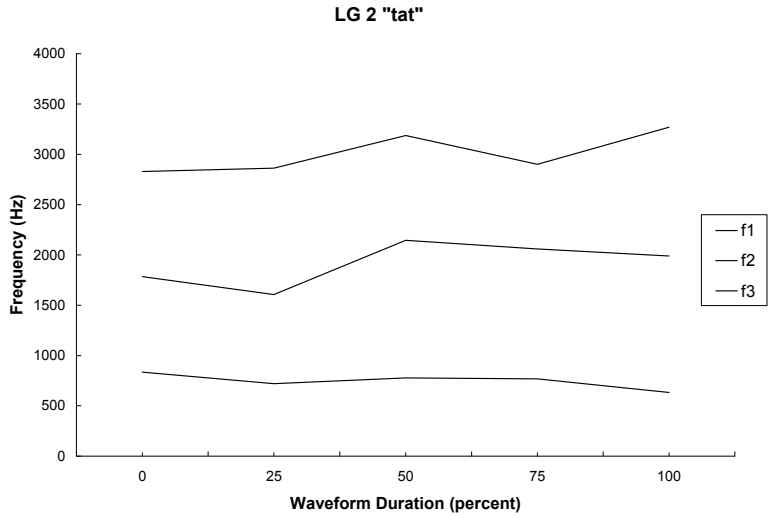


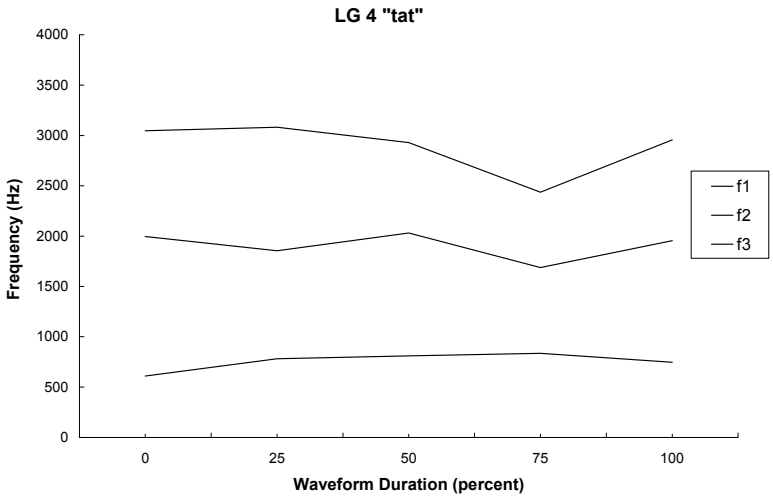




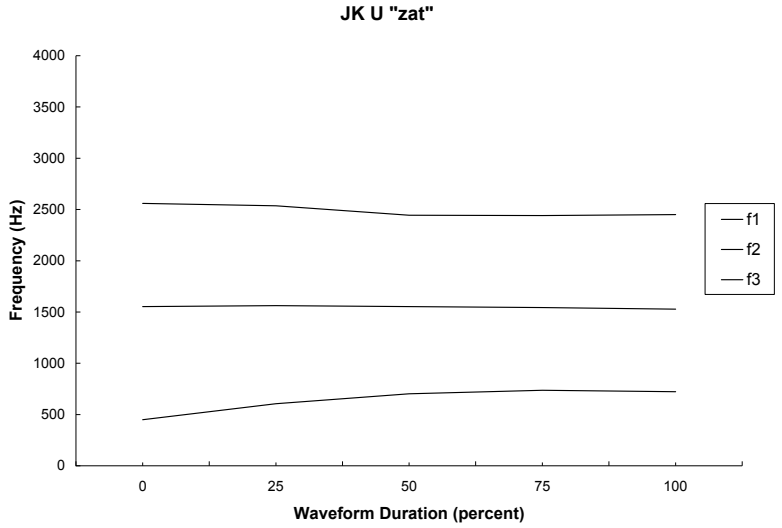
Formant Tracks for Talker LG producing the stimulus “tat”

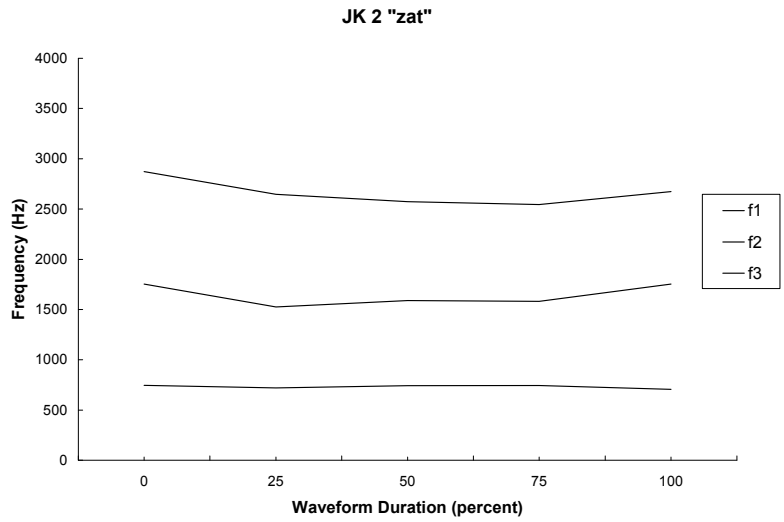


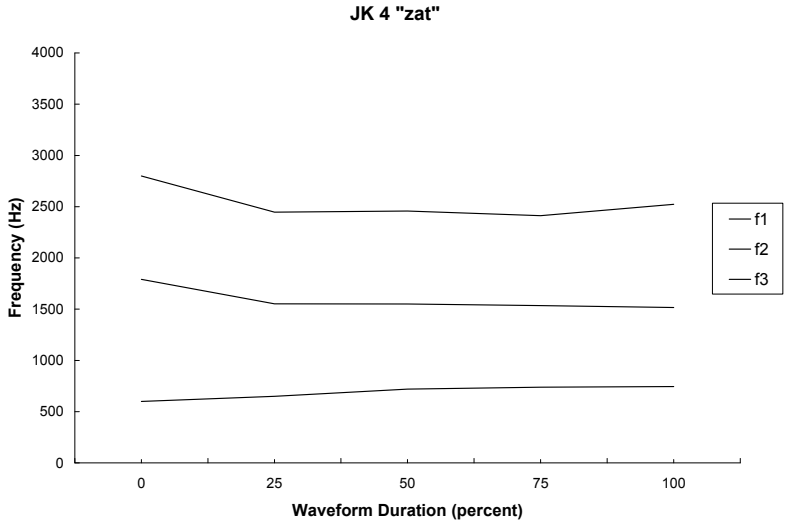




Formant Tracks for Talker JK producing the stimulus “zat”







Formant Tracks for Talker LG producing the stimulus “zat”